NASA TECHNICAL TRANSLATION NASA TT F-15,677

SPACE RESEARCH APPARATUS

Yu. K. Khodarev, L. I. Shatrovskiy,
V. V. Andreyanov, B. N. Rodionov,
P. Ye. El'yasberg, V. S. Etkin

Translation of "Apparatura dlya kosmicheskikh issledovaniy", "Nauka" Press, Moscow, 1973, 118 pages.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
WASHINGTON, D.C. 20546 AUGUST 1974

STANDARD TITLE PAGE

1. Report No. NASA TT F-15,677	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle SPACE RESEARCH APPARATUS		5. Report Date August 1974 6. Performing Organization Code
7. Author(s) Yu. K. Khodarev, L. I. Shatrovskiy, V. V. Andreyanov, B. N. Rodionov, P. Ye. El'yasberg, V. S. Etkin		8. Performing Organization Report No. 10. Work Unit No. 11. Contract or Grant No. NASw-2483
9. Performing Organization Name and Address SCITRAN Box 5456 Santa Barbara, CA 93108		13. Type of Report and Period Covered Translation
12. Sponsoring Agency Name and Address National Aeronautics and Space Administration Washington, D.C. 20546		14. Sponsoring Agency Code

15. Supplementary Notes

Translation of "Apparatura dlya kosmicheskikh issledovaniy", "Nauka" Press, Moscow, 1973, 118 pages.

16. Abstract

The articles of this collection encompass a broad range of questions associated with theoretical analysis and design of equipment used in conducting space experiments. The information on theoretical analysis of the possible apparatus solutions used to generate information streams aboard spacecraft is covered most completely. The articles on coding methods reflect the urgent necessity for more sophisticated onboard processing of the information obtained. The problems of spacecraft antenna testing, radiometric equipmen., and so on are examined.

The volume will be of interest to specialists connected with the design and construction of radioelectronic and radiophysical space equipment.

17. Key Words (Selected by Author(s))	18. Distribution Statement Unclassified - Unlimited		
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 189	22. Price

ANNOTATION

/2*

The articles of this collection encompass a broad range of questions associated with theoretical analysis and design of equipment used in conducting space experiments. The information on theoretical analysis of the possible apparatus solutions used to generate information streams aboard spacecraft is covered most completely. The articles on coding methods reflect the urgent necessity for more sophisticated onboard processing of the information obtained. The problems of spacecraft antenna testing, radiometric equipment, and so on are examined.

The volume will be of interest to specialists connected with the design and construction of radioelectronic and radiophysical space equipment.

TABLE OF CONTENTS

SERIES-LENGTH CODING UNDER A PRIORI UNCERTAINTY

Y. M. Shtar'kov and V. F. Babkin

ABSTRACT. A modification of the method of series length coding of bounded length is proposed. It is shown that if the binary symbol occurrence probabilities differ markedly, the proposed method may improve the message compression coefficient. The improvement depends on the degree of difference of the probabilities.

INTRODUCTION

/3*

The problem of the most effective statistical coding of a sequence of independent random events with known probabilities of their occurrence was solved long ago [1]. However, practical application of the methods developed has been hindered by the necessity for cumbersome "coding" and "decoding" tables and also absence of exact knowledge of the message individual probabilities.

If the source entropy $H < 1$, the first obstacle can be overcome with the aid of the statistical coding technique known as series-length coding. The first mention of this method is encountered in [2], a considerably more detailed analysis was made in [3], and a modification which is very convenient for practical purposes was proposed in [4, 5]. Several additional theoretical and experimental studies [6, 7] were made later. The attention devoted to this method is explained by the fact that this is a nonmnemonic technique and therefore there is no need to store coding and decoding tables. Moreover, in many cases the results obtained differ very little from the optimal results.

Considerably more serious is the problem of a priori uncertainty. Only a few studies have discussed particular statis-

* Numbers in margin indicate pagination in original foreign text.

tical coding methods with unknown individual message occurrence probabilities. These methods include, first of all, substitutions of the universal type [8]. A quite general approach to the solution of this problem was proposed quite recently, which unfortunately did not permit obtaining a coding technique acceptable from the viewpoint of complexity. The latter problem has been solved quite recently [9]. However, for many applications the proposed coding method is still not sufficiently simple. Therefore, in the present paper we reexamine series-length coding and propose modifications which are adapted for operation under conditions of partial or complete a priori uncertainty.

UNIFORM SERIES-LENGTH CODING OF FINITE LENGTH

Without significant loss of generality we limit ourselves to examining the sequence of symbols 0 and 1, which are statistically independent and have occurrence probabilities equal to q and $p = 1 - q$, respectively. We also assume that $q > p$.

In the original variant [2, 3] series-length coding consisted in breaking down the subject binary sequence into individual blocks ("enlarged symbols") of the form

$$\begin{matrix} 1 \\ 01 \\ 001 \\ 0001 \\ \cdots \end{matrix} \qquad (1)$$

and coding of each such block by a nonuniform code. However, /4 all the techniques for such transformation known at the present time are excessively complex. In [3] a ternary code is used which is by no means always convenient, while in [6] the use of codes with separator symbol is proposed, which is in principle nearly equivalent to the ternary code but leads to less simplicity of code combination formation. Moreover, examination of the infinite sequence of enlarged symbols (1) is impossible in practice. In order to find the next enlarged symbol it is sufficient to count

the number of zeros preceding the one. But any real counter can count up to a definite limit.

A different solution was proposed in [4, 5]. In place of (1) the system of enlarged symbols was examined

$$\begin{array}{ll} 1 & A_0, \\ 01 & A_1, \\ 001 & A_2 \end{array} \qquad (2)$$

where each succeeding block contains one more symbol 0 than the preceding block, the next-to-last and last blocks contain l symbols each, and the total number of blocks is thus equal to $l + 1$. Considering that (2) is a complete system of events, such partitioning is always possible.

At the same time, changeover to the finite system of enlarged symbols makes it possible to code using blocks of the same length, which is impossible in principle when using the system (1), which contains an infinite number of elements. Let the number r satisfy the condition

Then we can establish one-to-one correspondence between events from (2) and $l + 1$ (from -) by binary sequences of length r. It is simplest to match with any A_i the binary form of the number i (using all r places). Then when coding it is sufficient to count in an r-place counter the number of zeros prior to the occurrence of the first one (this rule changes somewhat for A_l).

For any fixed r it is well to select the group of events (2) as large as possible. Therefore we take l such that

$$\quad 1 \quad 2^r. \qquad (3)$$

The larger the ratio of the input block average length $\bar{n}$ to the output block average length $\bar{m}$, the higher is the effectiveness of statistical coding. In the present case the output block length is constant and equal to r. And it is not difficult to estimate the quantity $\bar{n}$, considering that the symbols 0 and 1 are statistically independent and the probabilities of the events A_0, A_1, A_2, ..., A_{l-1}, A_l are equal, respectively, to p, qp, q^2p, ..., $q^{l-1}p$, q^l:

$$\bar{n} = 1 \cdot p + 2qp + 3q^2p + \ldots + lq^{l-1}p + lq^l = \frac{1-q^l}{1-q} = \frac{1-q^{2^r}}{1-q}. \qquad (4)$$

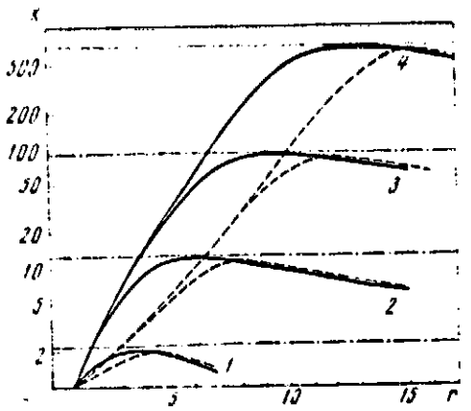
Then

$$K = \frac{\bar{n}}{m} = \frac{\bar{n}}{r} = \frac{1-q^{2^r-1}}{r(1-q)}. \qquad (5)$$

As is frequently done, we shall call the ratio $K = \bar{n}/\bar{m}$ the compression coefficient. Thus, K = K (q, r) and K = K (p, r). In the figure, the dashed curves show the dependence of K on r for p = 10^{-1}, 10^{-2}, 10^{-3} and 10^{-4}, respectively. Each curve has a maximum, and the value of the variable r for which the maximum is reached depends on p. The upper bound of the compression coefficient is /5

$$K^*(p) = \frac{1}{H(p)}, \qquad (6)$$

where $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the source entropy. In the figure, the values of K* (p) for selected values of the probability p are shown by the horizontal dash-dot lines. We see that $\max_r K(p, r)$ differs very little from K* (p) and the smaller p, the smaller this difference. Thus, for known p and q = 1 - p we can select a value of r such that the effectiveness of uniform series-length coding of finite length will differ very little from optimal statistical coding. This conclusion is valid for values of p < 10^{-1}. However, if p > 10^{-1} then even K* (p) < 2 and the advisability of any statistical coding becomes questionable.



Effectiveness of uniform and nonuniform coding for series:
1- $p = 10^{-1}$; 2- $p = 10^{-2}$; 3- $p = 10^{-3}$; 4- $p = 10^{-4}$

The fact that for different values of p the curves in the figure reach a maximum for different values of r = r (p) means that the subject method cannot be optimal for a wide range of p values. In selecting a definite value of r when constructing a coding scheme we must remember that the value of r which is optimal for one p may be far from optimal for another.

Generally speaking, there is nothing surprising in this. None of the statistical coding methods examined in [1, 6, 7] permit constructing a code which provides nearly optimal effectiveness for various values of p. Nor is uniform series-length coding of bounded length such a method. Only in [9-11] was the possibility shown and constructive techniques presented for constructing codes which are equally effective for various values of p. They differ only in slower approach to $K^*(p)$ with increase of $\bar{n}$.

From this viewpoint it is interesting to recall the results obtained with nonuniform coding of an infinite system of enlarged symbols (1). According to [2], when using a quaternary alphabet, which is equivalent to the use in binary code of a separating symbol consisting of two symbols [6], for any value of p in the interval from 0 to 0.5 the values of K (p) satisfy the condition $0.65\,K^*(p) \leq K(p) \leq 0.8\,K^*(p)$. Thus, at the cost of failure of K (p) to approach $K^*(p)$, a finite relative difference between K (p) and $K^*(p)$ is provided. The longer the length of the separating symbol (or the number of symbols in the code used), the

closer K (p) will be to K* (p) as $p \to 0$, but at the same time the smaller the value of K (p) as $p \to 1/2$.

For small p, when $K^*(p) \gg 1$, the difference between K (p) and K* (p) of a factor of 1.2-1.5 is quite acceptable in practice. However for p = 1/2 such coding would lead to an error of a factor of 1.5 in comparison with the basic binary sequence, which is completely unacceptable. It is obvious that with broadening of the range of p values of 1.0 the results deteriorate still more. /6 Finally, as we mentioned in the beginning of the section, examination of an infinite system of enlarged symbols, even as simple as (1), can be of theoretical interest only.

Let us return to examination of uniform series-length coding of finite length. Under conditions of partial a priori uncertainty we can use two different approaches to selection of the parameter r. Let the range of possible p values be given $p_1 \leq p \leq p_2$. In many cases it is important to ensure the largest value of the compression coefficient K in the worst possible case.

We see from the figure that for any r the compression coefficient takes the smallest value for $p = p_2$. Therefore we should select $r = \tilde{r}$, satisfying the condition $K(p_2, \tilde{r}) = \max_r K(p_2, r)$, which ensures a value of $K(p, \tilde{r}) \geq K(p_2, \tilde{r}) = \max_r K(p_2, r)$. For example, let $p \leq 10^{-2}$. In this case, according to Figure 1, $\tilde{r} = 8$ and $K(10^{-2}, 8) = 11.5$. For $p = 10^{-3}$ and $r = 8$ we obtain a larger value of K, equal to 28.2; however, this value now differs markedly from $K^*(10^{-3}) = 87.6$ and the smaller p, the larger will be K (p, r) and the larger its difference from K* (p). These examples show that uniform series-length coding of bounded length can be used under conditions of partial a priori uncertainty. Naturally, the result obtained will be better, the smaller the a priori uncertainty.

SIMPLE METHOD OF NONUNIFORM SERIES-LENGTH CODING OF BOUNDED LENGTH

We mentioned above that for known p we can select r such that uniform series-length coding (in the following we shall omit the words "of bounded length" for brevity, since the system (1) will not be examined further) yields a value of the compression coefficient which differs very little from $K^*(p)$. It appears that this was the reason why nonuniform series-length coding was not examined anywhere and was not mentioned. However, under a priori uncertainty conditions nonuniform coding is a natural technique which makes it possible to improve the results for a given range of p values. We emphasize immediately that with transition to nonuniform coding the extremely important property of simplicity, which makes it possible to avoid the use of special coding and decoding tables, must be retained.

In order that the nonuniform code improve the results for some range of p values, it is necessary that the $K(p, r)$ curves in the figure have "flatter" maxima. It is this feature that made it possible to select r such that $K(p, r)$ for various p differed very little from $\max_r K(p, r) \approx K^*(p)$. Naturally, this broadening of the maximum region must not (or may only to a slight degree) be accompanied by reduction of the magnitude of the maximum itself.

In this sense the problem becomes quite nontrivial. The only thing that can be said about conventional statistical coding of the system of enlarged symbols (2) for fixed p is that it brings $\max_r K(p, r)$ close to $K^*(p)$ precisely for this p. But this effect does not play any significant role, since $\max_r K(p, r) \approx K^*(p)$. As for the broadening of the maximum region, it cannot be completely guaranteed. Moreover, the suspicions that nonuniform statistical series-length coding for given p impairs

the characteristics of this nonuniform code in relation to other p seem completely justified. In this sense uniform coding behaves less selectively and from our viewpoint better than nonuniform coding for given p. Consequently the problem is to find those factors which influence the quite rapid falloff of the curves on both sides of their maxima, which are characteristic for different values of p, and which can be eliminated immediately for all p. In order to solve this problem we need to understand and clarify the behavior of the curves in the figure. /7

Let us examine the maximum of the K (p, r) curve for a fixed value of p. Since it differs very little from K* (p), uniform coding is close to optimal statistical coding. But this is valid if and only if the probabilities of the enlarged symbols A_0, A_1, ..., A_l do not differ markedly from one another.

Now let us begin to increase r (moving along the curve to the right). In this case l increases and the occurrence probability of A_l equal to q^l, A_{l-1} equal to $q^{l-1}p$, and so on decreases. At the same time the occurrence probabilities of A_0, A_1, A_2, and so on do not change. As a result, increase of r leads to ever greater and greater difference between the occurrence probabilities of the individual enlarged symbols and uniform coding becomes ever less and less optimal. This then explains the slow but steady decline of the K (p, r) curve with increase of r.

Thus, in order to eliminate or, more precisely, to weaken the decline of the K (p, r) curve with increase of r it is necessary to use a nonuniform code for which the length of the code combination corresponding to A_i is a monotonically increasing in the broad sense function of i. This is characteristic for any p, however the concrete form of such a function will depend on p. It is difficult to say anything definite about the results of such a transformation from the viewpoint of broadening the range

of p values. The only obvious effect is a shift of the maximum of each curve to the right. Moreover it is difficult to expect retention of coding and decoding simplicity.

Now let us examine the behavior of the curve to the left of the maximum. With decrease of r the occurrence probability of the enlarged symbol A_l approaches q, and all the remaining A_i ($i < l$) have practically the same probability. In this case we should code all A_i ($i < l$) by words of the same length and A_l by a shorter word. The simplest solution lies in coding the enlarged symbol A_l by a single binary symbol (0, for example) and any other symbol by r + 1 binary symbols, of which the first is equal to 1 and the remaining r coincide with the previously used words of the uniform code of length r. In this case the total number of enlarged symbols can be increased from $l + 1 = 2^r$ to $2^r + 1$, which is equivalent to transition from l to $l' = l + 1 = 2^r$. In this case the coding and decoding algorithms are just as simple as before. When coding, the combinations of $l' = l + 1$ zeros are replaced by a single zero and the series of $i < l'$ zeros (terminating in a one) are replaced by a one with ensuing r-place binary form of the number $0 \leq i \leq l' - 1 = 2^r$. Decoding proceeds just as simply.

We could replace A_l by a sequence of $r_0 = 2, 3, \ldots, r - 1$ symbols. When using lengths r + 1 for all the other code combinations this would permit increasing the total number of enlarged symbols to $2^r + 2^{r-1} + \ldots + 1$. This increase is always useful to some degree. On the other hand, however, for small values of r coding becomes less optimal than with the choice $r_0 = 1$. And for large r, when the A_l occurrence probability becomes small, selection of the quantity r_0 does not play any role. We can obviously state that the proposed nonuniform series-length coding variant leads to shift of the K (p, r) curve maxima to the left, towards smaller r. In fact, for uniform coding the

maximum is reached in the case when the A_l occurrence probability /8
is of the same order as that of A_0 ($\approx p$). But for the proposed nonuniform coding the maximum is reached with A_{l+1} occurrence probability equal to approximately 0.5. For $p < 0.5$ the second condition is always satisfied for smaller values of r than the first. This then means shift of the maximum in the direction of smaller r.

Now let us examine quantitatively the behavior of the compression coefficient $\tilde{K}(p, r)$ for nonuniform series-length coding. For greater convenience of comparison we immediately agree to take $\tilde{K}(p, r)$ to mean the compression coefficient for the case when $l' = 2^r$ and the length of all the code words other than that corresponding to A_{l+1} is equal to r + 1. This makes it possible to compare uniform and nonuniform coding with practically the same total number of enlarged symbols in (2) and the same maximal series length. Thus, $\tilde{K}(p, r) = \bar{n}/\bar{m}$. The quantity $\bar{n}$ can be found from (4) with replacement of l by $l' = l + 1 = 2^r$. On the other hand, $\bar{m} \neq r$. It is not difficult to verify that

$$\bar{m} = (1 - q^{l+1})(1 + r) + q^{l+1} = r(1 - q^{l+1}) + 1. \tag{7}$$

Now, using (4), (5), (7), we obtain

$$\tilde{K}(p, r) = \frac{1 - q^{l+1}}{1 - q} : [r(1 - q^{l+1}) + 1] = \frac{1}{(1 - q)\left[r + \frac{1}{1 - q^{l+1}}\right]} = \frac{1}{(1 - q)\left[r + \frac{1}{1 - q^{2^r}}\right]}. \tag{8}$$

If we neglect the difference between $1 - q^l$ and $1 - q^{l+1}$ we obtain the simple approximate formula

$$\tilde{K}(p, r) \approx \frac{K(p, r)}{1 - q^{2^r} + \frac{1}{r}}. \tag{9}$$

The solid curves in Figure 1 show the behavior of $\tilde{K}(p, r)$ as a function of r for the same probability values as before.

Comparison of the K (p, r) and $\tilde{K}$ (p, r) curves for the same value of p shows significant broadening of the $\tilde{K}$ (p, r) maximum region at the expense of marked increase of the compression coefficient for small r. Thus the posed problem is solved at least to some degree.

We note that, as expected, the $\tilde{K}$ (p, r) curve maxima are shifted to the left relative to the K (p, r) maxima. This means that for any problem formulation the best value of r is less than for uniform coding. This leads to decrease of the maximal series-length, which may be very useful in the presence of individual nonstationary segments in the sequence. It is very interesting that $\max_r \tilde{K}$ (p, r) > $\max_r$ K (p, r) for any p. In other words, broadening of the maxima region is accompanied not by reduction but rather even by some increase of the maximum itself. Finally, we note that in the region of large r the $\tilde{K}$ (p, r) values are somewhat lower than K (p, r), but this difference is not significant. In accordance with (10) this occurs for sufficiently large r, when $p^r < \frac{1}{2}$. However, the deterioration introduced in this case does not exceed 1/r and is small in comparison with the one in the denominator. Thus the proposed method makes it possible to obtain immediately several advantages. Over the entire range of r values the compression coefficient $\tilde{K}$ (p, r) either exceeds or practically coincides with K (p, r). The width of the $\tilde{K}$ (p, r) maximum region is greater than for K (p, r) and the magnitude of the maximum itself is larger. The maximum is reached for a smaller value of the parameter r and correspondingly smaller maximal series-length.

In conclusion we shall illustrate the noted advantages by examples. To this end we use the examples considered previously for $\tilde{K}$ (p, r) and then compare the results obtained. For example, if [illegible] and we are required to maximize the value of $\tilde{K}$ in the worst possible case, then in accordance with the figure we /9

find that we should take r = 6 and this provides a value of the compression coefficient no worse than K = 12.1. This result nearly coincides with that obtained previously, and this is quite natural, since the present problem reduces to determine $\max_r \tilde{K}(p^*, r)$, which as we noted previously differs very little from $\max_r K(p, r)$. However, for $p^* = 10^{-3}$ $\tilde{K}(10^{-3}, 6) = 44.4$, i.e., it is considerably larger than for the selection (which is optimal in this sense) of the parameter in the uniform coding case ($K(10^{-3}, 8) = 28.2$). The difference will be somewhat greater for $p = 10^{-4}$, $K(10^{-4}, 6) = 60.7 > K(10^{-4}, 8) = 31.5$. The examples show that in those cases when it is known in advance that the probability p will not exceed some quite small magnitude, nonuniform stress-length coding is better than uniform. However, if in reality the probability p is noticeably less than the indicated limit this difference may be very significant.

We shall list some factors which can influence the choice between uniform and nonuniform series-length coding under conditions of partial a priori uncertainty. First, the optimal series-lengths are different in the two cases. The optimal series-length is shorter for nonuniform coding, and therefore returning of the coder with the appearance of nonstationarities in the process being coded may be simplified considerably. Second, with optimal choice of the length of the series being coded uniform and nonuniform coding provide approximately the same gain, quite close to the limiting possible gain. In those cases when the a priori estimate of the occurrence probability of any symbol (0 or 1) is not sufficiently exact, and in reality this probability may be considerably less, nonuniform series-length coding may provide considerable improvement, which will be greater the more the true probability differs from its estimate.

REFERENCES

1. Fano, R.M. Peredacha informatsii. Statisticheskaya teoriya svyazi (Transmission of Information, Statistical Theory of Communication). Mir Press, Moscow, 1965.

2. Shannon, C.E. In the collection: Information Theory and Cybernetics. Foreign Literature Press, Moscow, 1963, p. 243.

3. Elias, P. In the collection: Information Theory and Its Applications. Fizmatgiz Press, 1959, p. 275.

4. Blokh, E.L. Elektrosvyaz. Vol. 1, 1959, p. 76.

5. Blokh, E.L. In the collection: Problemy Peredachi Informatsii (Information Transmission). Press of the Academy of Sciences of the USSR, No. 5, 1960.

6. Meshkovskiy, K.A. and N.Ye. Kirillov. Kodirovaniye v Tekhnike Svyazi (Coding in Communication Technology), Svyaz' Press, Moscow, 1966.

7. Novik, D.A. Effektivnoye Kodirovaniye (Effective Coding). Energiya Press, Moscow, 1965.

8. Laemel, A.E. Proc. Symp. Math. Theory of Automata, New York, 1962, p. 241.

9. Fitingof, B.F. In the collection: Problemy Peredachi Informatsii. Vol. 2, No. 2, 1966, p. 3.

10. Fitingof, B.M. In the collection: Problemy Peredachi Informatsii. Vol. 3, No. 3, 1967, p. 28.

11. Babkin, V.F. In the collection: Problemy Peredachi Informatsii. Vol. 7, No. 4, 1971, p. 13.

CODING OF DISCRETE MONOTONIC FUNCTIONS

A. B. Kryukov

ABSTRACT. Two techniques are proposed for numbering all monotone functions which permit economic coding of any such functions by a uniform code. We analyze the effectiveness of this function representation and, in particular, it is shown that for a fixed number of quantization levels the compression coefficient increases without limit with increase of the number of function readings. A characteristic feature of the techniques examined is the possibility of forming the code word in real time as the coded function readings arrive.

INTRODUCTION

/10

We examine a set of discrete functions, each of which is specified by a sequence of readings $\{s_1, s_2, \ldots, s_i, \ldots, s_n\}$, where the readings s_1 can take any integral values from 1 to m. The number of all such possible functions is equal to m^n; therefore, for coding any of them by a uniform code, we need to expend $\langle \log_q m^n \rangle$ q-nary units. (Here and hereafter $\langle \xi \rangle$ denotes the smallest integer larger than or equal to ξ.) In certain cases we can impose qualitative limitations on the nature of the function behavior. Then the number of possible functions reduces and, consequently, such a representation of the functions becomes redundant. For example, in a telemetry system with cyclic sampling of the sensors, as a result of the use of compression there is formed a sequence of significant readings whose addresses (in the limits of a single cycle) form a strictly increasing sequence. In [1] an effective technique for coding such address sequences is proposed, based on numbering all the strictly increasing functions.

In the present study, we propose two techniques for numbering functions from a broader class — functions which are monotonic

in the broad sense. An example of a case when such a limitation is imposed on the functions is the adaptive telemetry system in which the same sensor may be sampled several times in succession. In this case the significant reading addresses form a nondecreasing sequence. The techniques proposed below make it possible to encode economically by a uniform code nondecreasing or nonincreasing discrete functions and also both together. In the latter case it is necessary to expend one additional bit to indicate the sign of the increment (nondecrease or nonincrease).

In the following, for definiteness, we shall examine only nondecreasing functions $s_1 \leqslant s_2 \leqslant \ldots \leqslant s_i \leqslant \ldots \leqslant s$.

CODING TECHNIQUES

Technique 1. In this case, coding is accomplished in two steps. In the first step, the function is coded by a binary sequence of length m + n - 1 and in the second step, the obtained sequence is recoded into a different, more compact binary sequence or number.

We assign to each nondecreasing function a sequence of zeros and ones $a_1, a_2, \ldots, a_{m+n-1}$. This sequence consists of two parts. The first n-1 sequence elements are formed using the rule

$$a_i = \begin{cases} 1, & \text{if } s_{i+1} = s_i, \\ 0, & \text{if } s_{i+1} > s_i, \end{cases} \quad i = 1, 2, \ldots, n-1,$$

and the remaining m elements are formed using the rule

$$a_{n-1+j} = \begin{cases} 1, \text{ if } & j \in S, \\ 0, \text{ if } & j \in S, \end{cases}^{*} \quad j = 1, 2, \ldots, m,$$

where S is the set of values which the function being coded takes.

The formation of this sequence is shown in Figure 1. The first n-1 sequence elements are written in row form, so that the i-th element of this part of the sequence corresponds to the (i+1)-th function reading. The remaining m sequence elements /11

* Translator's note: There is obviously a mistake in the equation in the foreign text.

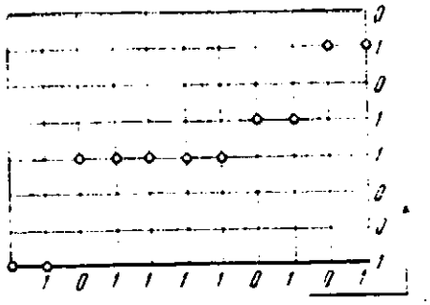


Figure 1. Formation of binary sequence of length m+n-1 (m=8, n=11)

are written in column form, so that the (n-1+j)-th element corresponds to the j-th quantization level. The order of sequence element numbering in the figure is shown by the arrow. In the row, the symbol 1 denotes those readings which do not differ from the preceding, while in the column the symbol 1 denotes the values which the function takes. Thus, to each nondecreasing function there corresponds a binary sequence of length m+n-1, and to the different functions there correspond different sequences which differ in the location of the ones. We note that a characteristic property of these sequences is the fact that the number of ones in the sequence is always constant and equal to n; the number of zeros is correspondingly equal to m-1. In fact, to the first reading there always corresponds the symbol 1 in the column and each following reading either does not differ from the preceding (then a symbol 1 in the row corresponds to this reading) or it takes a new value (then a symbol 1 in the column corresponds to it). Proof of one-to-one correspondence between the functions being coded and the sequences may be based on the fact that the positions of the zeros in the row and ones in the column are simply the coordinates of the characteristic points (jumps) of the function being coded.

In the second step recoding of the obtained sequence into a shorter binary sequence (or number) is accomplished. The number of binary sequences of length m+n-1 containing n ones is equal to C^n_{m+n-1}. They can be recoded by a special technique [1] so that $\langle \log_q C^n_{m+n-1} \rangle$ q-nary units will be expended on each. To

this end, each sequence is assigned a definite number A, where $0 \leqslant A \leqslant C^n_{m+n-1} - 1$. For formation of the numbers A, we use the numbers of those sequence positions where ones (or zeros) appear. We note that, generally speaking, the numbers A can be formed by both ones and zeros of the sequence, since $C^n_{m+n-1} = C^{m-1}_{m+n-1}$, but the number of arithmetic operations is proportional to the number of those symbols which are used for this purpose. Thus the number of arithmetic operations is proportional to min (n, m-1).

Technique 2. While in the first technique each function was initially assigned a binary sequence of length m+n-1 and then this sequence was coded, in the second technique direct coding of the functions is accomplished, bypassing the binary sequence formation stage. To this end we number the functions being coded, i.e., each of them is assigned a definite number B, where $0 \leqslant B \leqslant C^n_{m+n-1} - 1$. Naturally, the numbers B corresponding to different functions are different. The rule for forming these numbers is as follows.

As before, we assume the function is given by the sequence of readings $\{s_1, s_2, \ldots, s_i, \ldots, s_n\}$. Then the number B for this function is defined as the sum /12

$$B = \sum_{i=1}^{n} C^{i}_{s_i+i-2}, \qquad (1)$$

where C^{β}_{α} is assumed equal to zero if $\beta > \alpha$. It is not difficult to show with the aid of relation $C^{\beta}_{\alpha} = C^{\beta}_{\alpha-1} + C^{\beta-1}_{\alpha-1} + \ldots + C^{0}_{\alpha-\beta+1}$ that in accordance with (1) different numbers B correspond to the different functions.

Let there be given two nondecreasing functions $\{s'_1, s'_2, \ldots, s'_i, \ldots, s'_n\}$ and $\{s''_1, s''_2, \ldots, s''_i, \ldots, s''_n\}$, and let the largest number of the reading where the functions differ be equal to i, and all the

readings with numbers larger than i be the same for both functions, i.e., $s'_{i+1} = s''_{i+1}, s'_{i+2} = s''_{i+2}, \ldots, s'_n = s''_n$. For definiteness, we take $s'_i > s''_i$ and show that in this case $B' > B''$ always.

The terms in (1) corresponding to readings with numbers larger than i are the same for both functions: we denote their sum by R. Since readings which are larger in magnitude correspond to larger terms in (1), the "worst" case will be that in which $s'_1 = s'_2 = \ldots = s'_{i-1} = 1$, and $s''_1 = s''_2 = \ldots = s''_i = s'_i - 1$. Then the first i-1 terms in (1) for the function $\{s'_i\}$ are equal to zero and

$$B' = C^i_{s'_i+i-2} + R.$$

For the function $\{s''_i\}$

$$B'' = C^1_{(s'_i+i-2)-i} + C^2_{(s'_i+i-2)-(i-1)} + \ldots + C^i_{(s'_i+i-2)-1} + R =$$
$$= C^i_{s'_i+i-2} - 1 + R = B' - 1.$$

Thus, B' > B" even in the worst case, and this inequality is satisfied for any i (i = 1, 2, ..., n).

The decoding rule amounts to the following. In the fact step, we seek that value of k (k = 1, 2, ..., m) for which the inequalities are satisfied

$$C^n_{k+n-2} \leqslant B < C^n_{k+n-1}. \qquad (2)$$

Let (2) be satisfied for $k = s^*_n$. Then in the next step we seek that k (k = 1, 2, ..., s^*_n) for which the inequalities are satisfied

$$C^{n-1}_{k+(n-1)-2} \leqslant B - C^n_{s^*_n+n-2} < C^{n-1}_{k+(n-1)-1}. \qquad (3)$$

Let the condition (3) be satisfied for the value $k = s^*_{n-1}$. [In these inequalities, as in (1), C^β_α is assumed equal to 0 if $\beta > \alpha$.] This procedure is continued until obtaining the sequence $\{s^*_1, s^*_2, \ldots, s^*_n\}$. We note that if at any stage of the decoding procedure

— for example, when seeking the value of the r-th reading — the corresponding inequality becomes an equality for $k = s_r^*$, the procedure may be terminated, since all s_i^* having $i<r$ are equal to 1. Uniqueness of decoding is easily proved with the aid of the relation $C_a^b = C_{a-1}^b + C_{a-1}^{b-1} + \ldots + C_{a-b+1}^0$. By virtue of one-to-one correspondence between the numbers B and the functions being coded, $\{s_i^*\} = \{s_i\}$.

The coding and decoding procedures described above can be illustrated with the aid of a special table composed from the numbers C_{n-i-2}^k ($i = 1, 2, \ldots, n$; $k = 1, 2, \ldots, m$): /13

m	C_{n-1}^1	C_n^2	C_{n+1}^3	C_{n+2}^4	$\ldots$	C_{n+m-2}^m
k	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
4	$C_3^1 = 3$	$C_4^2 = 6$	$C_5^3 = 10$	$\boxed{C_6^4 = 15}$	$\ldots$	C_{n+2}^m
3	$C_2^1 = 2$	$\boxed{C_3^2 = 3}$	$\boxed{C_4^3 = 4}$	$C_5^4 = 5$	$\ldots$	C_{n+1}^m
2	$C_1^1 = 1$	$C_2^2 = 1$	$C_3^3 = 1$	$C_4^4 = 1$	$\ldots$	C_n^m
1	$\boxed{0}$	0	0	0	$\ldots$	0
	1	2	3	4	i	n

With the exception of the bottom row, consisting of zeros only, this table is a Pascal triangle in which the extreme left column of numbers $C_0^0, C_1^0, \ldots, C_{m-1}^0$ is missing. Like the Pascal triangle, this table can be constructed recursively with the aid of the relation $C_a^b = C_{a-1}^b + C_{a-1}^{b-1}$.

Let us examine coding and decoding with the aid of this table for an example. Let m = 4, n = 4 and the function be given by the readings {1, 3, 3, 4}. By "superposing" the function on the table, we find the number B as the sum of the table elements corresponding to the function values: B = 0 + 3 + 4 + 15 = 22. (In the table, these elements are shown bracketed.) This result naturally coincides with the value of B calculated using (1). We see from the table that the minimal value of B is equal to 0 for the function {1, 1, ..., 1}, while the maximal value for

{m, m, ..., m} is equal to

$$B_{max} = C^1_{m+1} + C^2_m + \ldots + C^{n}_{(m+n-1)-1} + C^n_{m+n-1} + 1.$$

Now let us perform the decoding. In the first step, we seek in the fourth column the largest number not exceeding B=22. This is the number 15 located in the fourth row. The row number is equal to the reading value, therefore $s_4 = 4$. We calculate the difference 22-15 = 7 and in the third column we seek the largest number not exceeding 7. This is the number 4, located in the third row, which means that $s_3 = 3$. We then calculate the difference 7-4 = 3, seek in the second column the largest number not exceeding 3, and find in the third row of this column the number 3. This means that $s_2 = 3$ and $s_1 = 1$, since 3-3 = 0. Thus, as a result of decoding we obtain the original reading sequence {1, 3, 3, 4}.

We note that, when forming the numbers B, the number of arithmetic operations is determined only by the value of n. Therefore, when n > m - 1, a smaller number of arithmetic operations is required for realization of the first technique. However, at the same time coding using the first technique is conducted in two steps, which creates additional difficulties not characteristic of the second technique. It is significant that both techniques permit coding of the functions directly at the rate of reading arrival, since formation of the numbers A and B can be accomplished as the readings arrive and terminates with arrival of the last reading. In this case, storage of the preceding reading values is not required.

ESTIMATE OF EFFECTIVENESS

The examined techniques for coding nondecreasing functions make it possible to expend $\langle \log_q C^n_{m+n-1} \rangle$ q-nary units on each of them. In the conventional coding technique, this requires

$n\langle\log_q m\rangle$ q-nary units. For $m,\ n \geqslant 2\ C^n_{m+n-1} < m^n$, therefore the proposed coding techniques make it possible to reduce the number of symbols expended. The magnitude of the advantage obtained in this case, i.e., the degree of reduction of the number of q-nary units, is usually characterized by the compression coefficient /14

$$K_{com} = \frac{n\langle\log_q m\rangle}{\langle\log_q C^n_{m+n-1}\rangle}. \qquad (4)$$

The magnitude of the compression coefficient (4) depends significantly on m and n. In order to clarify the nature of this dependence, we shall examine two cases.

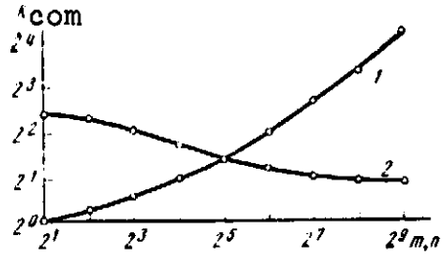


Figure 2. Compression coefficient as function of n and m (q=2). 1- m=32, n variable; 2- n=32, m variable (smooth curves are drawn through the calculated points for better visualization)

Case 1. m fixed, n variable. Figure 2 shows the compression coefficients for various n ($n = 2^1, 2^2, \ldots, 2^9$) for the particular case m = 32 (curve 1). We see that the compression coefficient increases monotonically with increase of n. As $n \to \infty$, neglecting the roundoff symbol <•> and using the relations

$$\ln(1+\varepsilon) = \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3} - \ldots,$$
$$\frac{1}{1-\varepsilon} = 1 + \varepsilon + \varepsilon^2 + \ldots,$$

we can transform (4) to the form

$$K_{com} \simeq \frac{n\log_q m}{(m-1)\log_q n}\left[1 + O\left(\frac{1}{\log_q n}\right)\right], \qquad (5)$$

where $O(\varepsilon)$ denotes a function of ε such that

$$\lim_{\varepsilon \to 0} \frac{O(\varepsilon)}{\varepsilon} = \text{const.}$$

It follows from (5) that the compression coefficient increases without bound as $n \to \infty$. Thus, the larger n, the greater the advantage obtained using the proposed coding techniques.

Case 2. n fixed, m variable. This case is examined to clarify the dependence of K_{com} on the number of quantization levels. The values of the compression coefficient for various m ($m = 2^1, 2^2, \ldots, 2^9$) and fixed $n = 32$ are also shown in Figure 2 (curve 2). For fixed n and $m \to \infty$, the expression (4) with the aid of the same transformations as in the preceding case can be reduced to the form

$$K_{com} \simeq 1 + O\left(\frac{1}{\log_q m}\right). \qquad (6)$$

In this case, in contrast with the preceding case, the expression for the compression coefficient has the limit

$$\lim_{\substack{m \to \infty \\ n=\text{const}}} K_{com} = 1.$$

The difference in the results obtained for cases 1 and 2 is explained as follows. The number of nondecreasing functions (and the number of symbols expended for nonredundant coding) depends on n and m practically identically. (These relations are absolutely identical if we examine m-1 rather than m as the argument.) At the same time, the basic representation of the functions, i.e., conventional coding with the aid of $n \langle \log_q m \rangle$ q-nary units, does not have such symmetry — namely, the number of symbols expended increases far faster with increase of n than with increase of m. In other words, for n>m this form of basic monotone function representation has greater redundancy than for m>n. Therefore for $m \gg n$ nonredundant coding of nonde- /15

creasing functions may not lead to any appreciable advantage in comparison with the conventional coding technique, since the redundancy of the latter is not large in this case.

This last remark makes it possible to conclude that non-redundant coding (numbering) of histograms by a uniform code may not lead to any significant reduction of the number of symbols expended. In fact, we can construct a distribution function (integral law) from each histogram. Since the distribution functions do not decrease, one of the techniques proposed above can be used for their coding. However, this does not lead to any marked advantage, since the relation $m \gg n$ is characteristic for histograms.

REFERENCES

1. Lynch, T.J. Proceedings of JIEEE (Letters), Oct. 1966, p. 54.

2. Babkin, V.F. Chetvertyy simpozium po probleme izbytochnosti v informatsionnykh sistemakh (Fourth Symposium on Redundancy in Information Systems), Reports, Vol. 1, Leningrad, 1970.

SIMPLE METHOD OF JUMBERING BINARY SEQUENCES WITH GIVEN NUMBER OF UNITS

Yu.M. Shtar'kov and V.F. Babkin

ABSTRACT. A method for coding messages with a priori unknown occurrence probabilities is proposed which has a very simple realization. The calculations presented show that the method effectiveness is close to the limiting possible value — the effectiveness of universal coding. Various modifications of the method are discussed.

INTRODUCTION

In direct application of the statistical coding ideas, it is necessary to use cumbersome tables, whose complexity increases exponentially with increase of the length n of the sequences (blocks) being coded. However, in many cases coding can be reduced to uniform numbering of all blocks with given limitations (properties). This approach was first used in [1] to code an independent source with unknown statistics. In this case, the technique proposed in [2] was used for uniform numbering of all binary symbol blocks containing a fixed number t of ones. If $i_1, i_2, \ldots, i_t$ are the positions of ones in a specific block, it is assigned the number

$$0 \leqslant a_t = C_{i_1-1}^1 + C_{i_2-1}^2 + \ldots + C_{i_t-1}^t \leqslant C_n^t - 1, \qquad (1)$$

where C_i^j is considered equal to zero if $i = j - 1$. This numbering ensures one-to-one correspondence of all the a_t and blocks of length n with t ones.

The labor involved in the numbering (1) increases only as a power-law function of n. Nevertheless, a specialized device /16 which computes or stores the binomial coefficient values would be quite complex. Therefore, the problem arises of finding numbering techniques whose effectiveness is nearly optimal and whose realization is considerably simpler. Since the uniform numbering problem arises only in application to equi-probable

sequences, in the general case the effectiveness can be defined as the average number of symbols expended in describing a single number. It is precisely on the basis of this criterion that we evaluate the effectiveness of the simple numbering technique examined below.

NUMBERING ALGORITHM

The only simple technique for statistical coding of binary sequences is known as series-length coding (its description and a detailed bibliography are presented in [3]). The basic sequence is broken down into "enlarged symbols" A_i, containing i zeros and a following one ($0 \leq i < l$, $l = 2^r - 1$) or $l = 2^r - 1$ zeros (A_l). Then each A_i is replaced by an r-digit number B_i, whose magnitude is equal to i. For any value of the probability $p \leq 1/2$ of one occuring, we can select the parameter r to ensure high coding effectiveness. This method is realized very simply in practice with the aid of an r-place binary counter. We use series-length coding to number the blocks containing t ones. Beginning with the first symbol of the block, we partition into enlarged symbols A_i up to and including the last, t-th, one (this is always possible) and replace A_i by B_i. The obtained code words satisfy the following conditions:

1) a single code word corresponds to each block;
2) a unique block corresponds to each code word;
3) the ensemble of all code words has the prefix property [4].

The first assertion follows directly from the coding technique described above. For proof of the second assertion, it is sufficient to consider that each code word consists of r-place numbers B_i. Therefore, breaking any code word into groups of r symbols each, we define the sequence of numbers B_i. Replacing each of them by the corresponding enlarged symbol A_i, we restore uniquely the initial symbols of the original block up to and

including the t-th one. Adding to the restored part of the block a series of zeros complementing the already obtained number of symbols to n, we obtain the original block, which proves the second assertion.

Finally, the prefix property means that no code word is the beginning of any other code word. In the present case, this property is due to the fact that each code word contains t "numbers" B_i, $i \neq l$ and $s \leqslant (n-t)!$ numbers B_l, and the last number is not equal to B_l (coding terminates after the appearance of the t-th one and identification of the corresponding $A_i \neq A_l$). In fact, let us assume that the beginning of the code word coincides with another code word. Then, just as the second code word, this beginning must contain t numbers B_i ($i \neq l$). But since the last number of any code word is not equal to B_l, on the whole the first code word must contain no less than t + 1 numbers B_i ($i \neq l$), which contradicts the code word structure. Therefore, the beginning of any code word cannot coincide with another code word, which proves the prefix property.

In accordance with the first property, any block or sequence of blocks is uniquely transformed into a code word sequence. Conversely, according to the third property, from any sequence of symbols we can identify uniquely the ensemble of any known number of code words if the position of the first symbol of the first code word is known. According to the second property, from the /17 identified code words we can recover the original sequence of blocks. Thus, there exists a one-to-one correspondence between the sequence of blocks and the sequence of code words. This makes it possible to consider the code words as numbers of the corresponding blocks. The fundamental difference between this numbering technique and (1) consists in the fact that the number "length" m_t is not a constant quantity. In accordance with the criterion formulated above, the effectiveness is determined by

comparing $\log_2 C_n^t$ (the minimal fixed number length) and the average number $\bar{m}_t$ of symbols used to describe a single number when using the subject technique. Since all C_n^t blocks with t ones are equally probable, then

$$\bar{m}_t = \frac{1}{C_n^t} \sum m_i, \qquad (2)$$

where the summation extends over all such blocks.

It is obvious that $\bar{m}_t = \bar{m}_t(r)$. Therefore, for each t we must use the coding parameter $r = r_t$ which minimizes $\bar{m}_t$. Since series (of zeros) length coding is effective for $p \leq 1/2$, for $t > n/2$ we obtain $r_t = 1$ and $\bar{m}_t \approx n$. In order to eliminate this drawback, for $t > n/2$ we will code not the original block but rather a "complementing" block, obtained by replacing the zeros by ones and the ones by zeros. It is obvious that then $r_t = r_{n-t}$ and $\bar{m}_t = \bar{m}_{n-t}$. Correspondingly, after decoding it is necessary to replace the obtained block by the complementary block.

Thus, in order to select r_t $(0 \leq t \leq n/2)$ and evaluate the effectiveness of the subject numbering technique we need to obtain the expression for $\bar{m}_t(r)$.

AVERAGE CODE WORD LENGTH

Let us determine the number $t + s_t$ of enlarged symbols obtained when using the algorithm described above, where $s_t = \frac{n-t}{l}$ is the number of enlarged symbols A_l. If $i_1, i_2, \ldots, i_t$ are the positions of the ones in a specific block and

$$a_k = i_k - i_{k-1} - 1 = s_i^{(k)} l + u_k, \quad 0 \leq u_k < l, \quad k = 1, 2, \ldots, t \qquad (3)$$

(i_0 is assumed to be equal to zero), then for the subject block

$$s_t = \sum_{k=1}^{t} s_i^{(k)}. \qquad (4)$$

In fact, it is not difficult to see that upon partitioning there will be obtained the following sequence of enlarged symbols

$$\underbrace{A_i A_i \ldots A_i}_{s_t^{(1)}} A_{u_1} \underbrace{A_i A_i \ldots A_i}_{s_t^{(2)}} A_{u_2} \ldots \underbrace{A_i A_i \ldots A_i}_{s_t^{(t)}} A_{u_t}.$$

And since $m_t = r\ (i + s_t)$, using (4) we obtain

$$\bar{m}_t = r(t - \bar{s}_t) = r\left(t + \sum_{k=1}^{t} \bar{s}_t^{(k)}\right), \tag{5}$$

where the overbar denotes ensemble averaging.

In accordance with (3)

$$P(s_t^{(k)} = y) = \sum_{x=yl}^{(y+1)l-1} P(\alpha_k = x), \tag{6}$$

where $P\ (s_t^{(k)} = y)$ and $P\ (\alpha_k = x)$ are the probabilities that $s_t^{(k)} = y$ and $\alpha_k = x$. Then, considering that $P\ (\alpha_k = x > n - t) =$ /18 0, we obtain

$$\bar{s}_t^{(k)} = \sum_{y=1}^{\infty} P(s_t^{(k)} = y)\, y = \sum_{y=1}^{\theta} P(s_t^{(k)} = y)\, y =$$
$$= \sum_{y=1}^{\theta-1} \left\{ \sum_{x=yl}^{(y+1)l-1} P(\alpha_k = x) \right\} y + \theta \sum_{x=\theta l}^{n-t} P(\alpha_k = x),$$

where $\theta = \left\lfloor \frac{n-t}{l} \right\rfloor$ and $\lfloor z \rfloor$ is the whole part of z. Thus the problem reduces to determining the distributions $P\ (\alpha_k = x)$, $k = 1, 2, \ldots, t$. Since all the blocks are considered to be equiprobable, $P\ (\alpha_k = x)$ is equal to the number of positions in which $\alpha_k = x$, divided by C_n^t.

We first of all show that

$$P(\alpha_k = x) = P(\alpha_1 = x),\ k = 2, 3, \ldots, t. \tag{7}$$

For this, we examine an arbitrary arrangement of t ones specified by the vector $(i_1, i_2, \ldots, i_t)$ or, what is the same, $(\alpha_1, \alpha_2, \ldots, \alpha_t)$. If $\alpha_k \neq \alpha_1$ there exists exactly one vector $(\alpha_1', \alpha_2', \ldots, \alpha_t')$

in which $\alpha_1^* = \alpha_k$, $\alpha_k^* = \alpha_j = \alpha_j$ $(j \neq 1, k)$. But then the number of arrangements in which $\alpha_1 = x$ is exactly equal to the number of arrangements in which $\alpha_k = x$. In fact, any arrangement with $\alpha_1 = \alpha_k = x$ is considered simultaneously in $P(\alpha_1 = x)$ and $P(\alpha_k = x)$ and to any arrangement with $\alpha_1 = x \neq \alpha_k$ there corresponds one and only one arrangement with $\alpha_1^* = \alpha_k$, $\alpha_k^* = x$ and $\alpha_j = \alpha_j^*$ $(j \neq 1, k)$. The equality of the number of arrangements with $\alpha_1 = x$ and $\alpha_k = x$ proves (7), and it is necessary to determine only $P(\alpha_1 = x)$. If the first one in the block is located at the $i_1 = \alpha_1 + 1$-th position, the remaining $t - 1$ ones may be arranged among the remaining $n - i_1$ symbols in $C_{n-i_1}^{t-1} = C_{n-x-1}^{t-1}$ ways. This expression defines the number of arrangements with the specified α_1 and $P(\alpha_1 = x) = C_{n-x-1}^{t-1}/C_n^t$. Then in accordance with (6)

$$\bar{s}_t^{(1)} = \frac{1}{C_n^t}\left[\sum_{y=1}^{\theta-1}\left\{\sum_{x=yl}^{(y+1)l-1} C_{n-x-1}^{t-1}\right\} y + \theta\left\{\sum_{x=\theta l}^{n-t} C_{n-x-1}^{t-1}\right\}\right], \quad \theta = \left\lfloor \frac{n-t}{t} \right\rfloor, \tag{8}$$

and from (5) and (7)

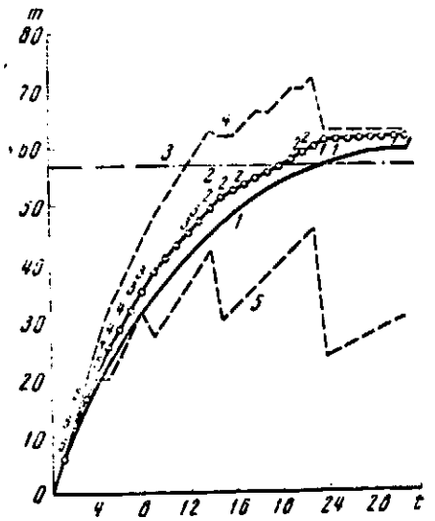
$$\bar{m}_t = r\left(t + \sum_{k=1}^{t} \bar{s}_t^{(k)}\right) = r\left(t + t\bar{s}_t^{(1)}\right) = rt\left(1 + \bar{s}_t^{(1)}\right). \tag{9}$$

Relations (8) and (9) define completely the average code word (block number) length.

EFFECTIVENESS OF THE SUBJECT NUMBERING METHOD

Theoretical analysis of (8) and (9) is not possible. Therefore, we calculated the quantities $\bar{m}_t$ for $n = 63$ and $t = 0.1, \ldots, 31$ (as noted above, $\bar{m}_t = \bar{m}_{n-t}$). Figure 1 shows the results of these calculations. Curve 1 shows the behavior of $\bar{m}_t^* = m_t^* = \log_2 C_n^t$, corresponding to optimal uniform numbering, and curve 2 shows the behavior of $\bar{m}_t$ when using the algorithm examined above. As we would expect, $m_t^* < \bar{m}_t$ for any $t > 1$, and $\max_t \{\bar{m}_t - m_t^*\} = \bar{m}_9 - m_9^* = \bar{m}_{15} - m_{15}^* = \bar{m}_{24} - m_{24}^* \simeq 4.2$ binary symbols, and $\max_t (\bar{m}_t - m_t^*)/m_t^* = (\bar{m}_4 - m_4^*)/m_4^* \simeq 1.28$. Thus, considerable simplification of the coding procedure is achieved at the cost of effectiveness deterioration /19

by no more than 4.2 binary symbols or a factor 1.28. In many cases such losses are quite acceptable. Thus, when coding a source with unknown statistics, the coding redundancy of a block of length n cannot be made less than $0.5 \log_2 n$[5], which in the present case constitutes 3.0 binary symbols and is comparable with the losses of this simple numbering technique.



Comparison of effectiveness of universal coding method and universal coding method with the use of series-length coding with block length n = 57

In applications such as space studies, certain other characteristics are also important — for example, the maximal possible gain in describing a block, equal to the difference of the maximal number length and the number of symbols in the block. In the optimal uniform numbering case, this difference is negative for any t (its maximum is reached for t = 31 and 32 and is equal to 3.3 binary symbols). When using the proposed numbering technique, $\bar{m}_t$ is also always smaller than n, but the maximal value of m_t may exceed $\bar{m}_t$ considerably and for large t may even exceed n. It is not difficult to see that

$$\max m_t = r_t\left(t + \left\lfloor \frac{n-t}{2^{r_t}-1} \right\rfloor\right), \qquad (10)$$
$$\min m_t = r_t t.$$

The horizontal dash-dot straight line 3 in the figure corresponds to the value n = 57, and the dashed broken lines 4 and 5 correspond to the upper and lower bounds of (10). The maximal value of m_t - n is reached for t = 23, and is equal to nine

binary symbols. This loss is not so large that the danger of obtaining a small group of "inconvenient" arrangement would force us to reject the use of this numbering technique.

There is a simple way to limit max $\{m_t - n\}$. We write ahead of each number an additional binary symbol and stipulate that if it is a zero the following number is constructed using the technique described above, and if it is a one the following symbols are the direct form of the "numbered" block. With this modification max $\{m_t - n\} = 1$ at the expense of increasing the length of most of the numbers by a single binary symbol.

Finally, we shall discuss briefly possible improvements of the subject method which permit approaching $\bar{m}_t$ to m_t^* and reducing the dispersion of m_t. The value of r_t is a function not only of the number t of ones but also of the block length n. For every n', where $1 < n' \leq n$, and every t', where $0 < t' < n'$, we can determine the optimal value of the parameter r : $r_t = r(n', t')$. Then the previously described numbering technique may be altered as follows. At the initial moment we use $r_t = r(n, t)$. Let the first enlarged symbol be A_i $(i < l)$. This means that in the remaining part of the block of length $n - i - 1$ symbols there are $t - 1$ ones. But for such a "shortened" block with $t - 1$ ones there exists the optimal value of $r_t = r(n - i - 1, t - 1)$, which may not coincide with $r(n, t)$. Therefore, the second enlarged symbol is identified and coded using the parameter $r(n - i - 1, t - 1)$. However, if the first enlarged symbol is equal to A_l, then in the second step of the numbering we use $r(n - l, t)$. After the /20 second step we again correct the parameter r, and so on until the appearance of the t-th one, after which the numbering process terminates. It is not difficult to see that in this case the prefix property is again satisfied and uniqueness of the decoding is ensured. The r (n, t) used in the first step is known a priori. Therefore we initially identify the r (n, t) first binary symbols

and the enlarged symbol A_1. In the next step we identify the $r(n - i - 1, t - 1)$ binary symbols if $i < l$ or $r(n - l, t)$ binary symbols if $i = l$. Using the identified binary symbols we determine the second enlarged symbol, the new value of r, and so on until the appearance of the t-th one.

The use of this procedure makes it possible to reduce $\bar{m}_t$ in comparison with the basic subject technique. This follows directly from the fact that in the basic subject numbering variant we select the value of r_t which is optimal in the mean for numbering the entire block (up to the t-th one), while in the proposed modification of the technique we select the optimal value of r in each step of the numbering process (after identifying the next enlarged symbol). The only drawback of this modification is the necessity for realizing the function r (n, t) of two variables. For r_t with fixed n this problem is resolved simply — it is sufficient to have a $\log_2 (n + 1)$-place counter with the same number of coincidence circuits and a $\log_2 \log_2 (n + 1)$-place counter, since r_t is a monotone function of t ($t \leq n/2$). In spite of the fact that r (n, t) is monotone with respect to each variable, its realization is incomparably more complex.

We note that it is possible to construct other modifications of the numbering method, for example, using nonuniform series-length coding [3] or modification associated with application of the subject technique in the universal coding procedure. However, these variants will not be examined here.

REFERENCES

1. Babkin, V.F. In the collection: Problemy Peredachi Informatsii. Vol. 7, No. 4, 1971, p. 13.

2. Lynch, T. Proceedings of IEEE. Vol. 54, 1966, p. 1490.

3. Shtar'kov, Yu.M. and V.F. Babkin. See article in present volumn.

4. Fano, R.M. Peredacha informatsii. Statisticheskaya teoriya svyazi (Transmission of Information. Statistical Theory of Communication). Mir Press, Moscow, 1965.

5. Krichevskiy, R.Ye. In the collection: Problemy Peredachi Informatsii. Vol. 4, No. 3, 1968, p. 48.

MULTIPURPOSE INFORMATION COLLECTION AND PROCESSING SYSTEMS

A. V. Kantor, S. M. Perevertkin and
T. S. Shcherbakova

ABSTRACT. We examine the construction and present a classification of multipurpose information collection and processing systems which are based on the use of information sorting with the aid of associative memories. Multiaddress compression, multichannel compression, and multipriority compression processes are described.

The problems of space telemetry impose urgent requirements for /20
the development of the so-called multipurpose collection and processing information systems. Here we mean by multipurpose data collection and processing systems, those which compress multichannel telemetry information with subsequent formation on board the spacecraft of several compressed data streams with prespecified characteristics. The requirement for development of systems of tnis sort is due, first of all, to the constantly increasing volume of measured information, increasing from experiment to experiment, which must be transmitted and recorded (with subsequent retransmission) /21
over a few channels with limited handling capacity, the necessity for integrating the telemetry system into the spacecraft control loop, and many other factors. In the present article, we consider

three possible forms of multipurpose information collection and processing system organization: multichannel, multiaddress, and multipriority.

Depending on the multipurpose system organization form adopted at their output, there are formed from the stream of non-equally-spaced significant multichannel telemetry information readings one or more streams of equally-spaced readings with one of the following characteristics:

1) in the case of multichannel organi ation of the multipurpose information collection and processing systems, the compressed information output streams are formed from those significant telemetered parameter readings which provide the specified values of the approximation error in each transmission channel; the number of output streams corresponds to the number of different values of the original analog parameter approximation error;

2) in the case of multiaddress organization of the multipurpose systems, the compressed information output streams are formed from the significant telemetered parameter readings with definite destination address; the number of streams is equal to the number of addressees;

3) in the case of multipriority organization, the significant reading sequence in the output stream is determined by the value of the priority labels assigned to each message source group; the output may be a single stream, or the formation of several output streams is possible.

Various combinations of the multipurpose systems listed above can be used to create quite flexible information collection and processing systems. We note that for creation of multipurpose telemetry information collection and processing systems based on the traditional adaptive discretization schemes, the number of parallel output stream formation channels would obviously be equal to the given number of these streams. Such multipurpose system construction

can lead to an inacceptably cumbersome scheme. In the multipurpose information gathering systems which use associative memories (AM), which will be examined hereafter, the formation of a whole series of compressed information streams is accomplished in a single channel, i.e., using the same equipment. Individual questions related with the possibilities of using AM in telemetry information collection and processing systems have been discussed in [1, 2].

The expanded block diagram of a multipurpose information collection and processing system using an AM is shown in Figure 1. The multipurpose telemetry information compression process is accomplished

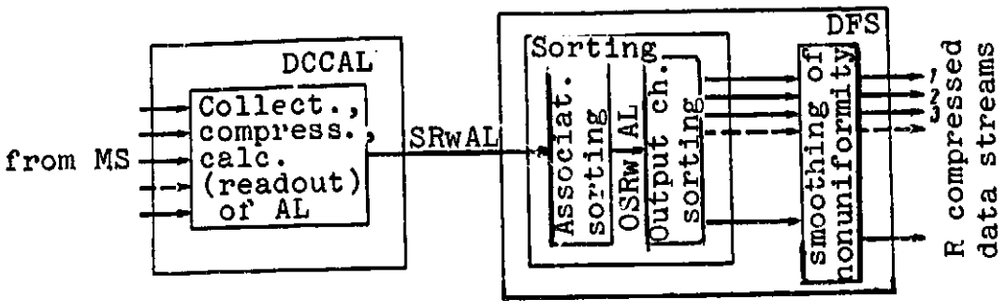


Figure 1. Block diagram of multipurpose information collection and processing system:

MS — message source; DCCAL — device for compressing and calculating associative labels; SRwAL — significant readings with associative labels; DFS — device for forming streams; OSR — ordered significant readings

sequentially by the device for collecting and compressing information and calculating (selecting) associative labels (DCCAL), and the device for forming compressed data output streams (DFS). We note that, in the case of multichannel organization of the multipurpose information collection systems, the telemetered parameter approximation error is the significant reading associative label, in the multiaddress case the destination address is the associative label, and in the multipriority case the message source priority rank is the associative symbol.

The output stream formation process includes both general associative sorting of the compressed telemetry information and the process of sorting the ordered significant readings with respect to the output channels with subsequent smoothing of the nonuniformities of these streams in a smoother. In the associative sorting process (Figure 2), there is formed a significant reading sequence, ordered in decreasing associative label order, termed hereafter the ordered significant reading table. Sorting of the ordered significant reading table with respect to the output channels leads to the formation /22

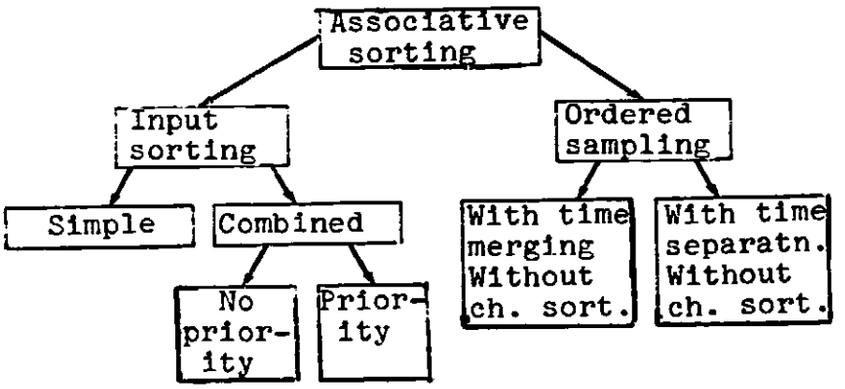


Figure 2. Associative sorting classification

of R (one for each output channel) ordered significant reading sequences, each of which includes readings with corresponding associative label values or ranges of values. In the following, these sequences are termed transmitted significant reading tables. The complete transmitted significant reading table includes, without exception, all readings from the general ordered significant reading table. The formation of the ordered significant reading table can be accomplished either in the process of entering the information into the AM, if an AM with input sorting is used, or in the process of reading the information from the AM, if an AM with ordered readout is used.

In the case of information sorting in the process of entering the readings into the AM with input sorting, we can speak of simple, combined, and combined-with-priority sorting of the readings. In

the case of simple information sorting, the readings are entered into the AM as follows:

1) in decreasing associative label order;

2) without one-to-one correspondence between the cell group address and the associative label values entered in any cell of this group;

3) with loss of readings corresponding to the minimal associative label values in the case of AM overflow.

In the case of combined information sorting, the significant readings are entered into the AM in the following sequence:

1) in decreasing order of the associative label values;

2) with one-to-one correspondence between the cell group addresses and the associative label values of the significant readings entered in any cell of this group;

3) with loss of readings with given associative label value if all the cells of the group intended for entry of readings with associative label values lying in the limits of the specified range are filled.

In the case of combined information sorting at the inlet of the AM with priorities, entry is accomplished:

1) in decreasing order of the associative label values;

2) with one-to-one correspondence between the cell group addresses and the associative label values of the readings entered in cells of this group, provided overflow of the cell group intended for entry of readings with large associative label value has not taken place; /23

3) with loss of readings with minimal associative label values.

Thus, in the simple information sorting case, we obtain floating distribution of the readings with respect to the associative label magnitudes in the associative memory. In the combined sorting case, we obtain a fixed distribution, and in the case of combined information sorting at the input to the AM with priorities, we obtain a mixed reading distribution which coincides with the fixed distribution in the absence of cell group overflow.

When using an AM with ordered readout, associative information sorting is accomplished in the process of readout of the readings from the memory in decreasing associative label value order. In this case, we can speak of quasi-fixed distribution of the significant readings stored in the associative memory in the process of ordered /24
significant reading table formating using the cyclic sampling method.

These associative sorting modes are shown in Figure 2. Figures 3, 4, 5 show schematically the ordered significant reading tables for the AM input or output associative sorting modes examined above.

The ordered significant reading table is represented in the /25
form of a column of rows. Each row corresponds to a significant reading with some particular associative label value, entered either in the corresponding AM cell (AM with input sorting) or in the AM output register (AM with ordered readout).

Significant reading sorting with respect to the output channels (Figure 6) for multipurpose information compression in the general case can be accomplished using the corresponding program in accordance with the selected output stream formation criterion. This sorting is based on combining, in each output stream, significant readings with different associative label values. In the case of formation of only a single stream of significant readings which follow in a definite sequence, we shall speak of the simplest

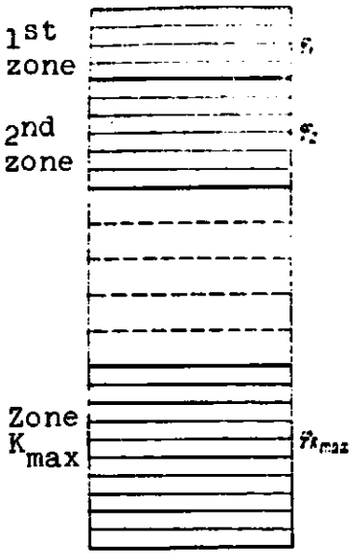


Figure 3. Table of ordered significant readings with fixed distribution:

$\phi_i = \phi_{[K_{max-(1-1)}]al}$ — table zone used for entering readings with associative label values; al — minimal possible associative label value

$1 \leq K \leq K_{max}$

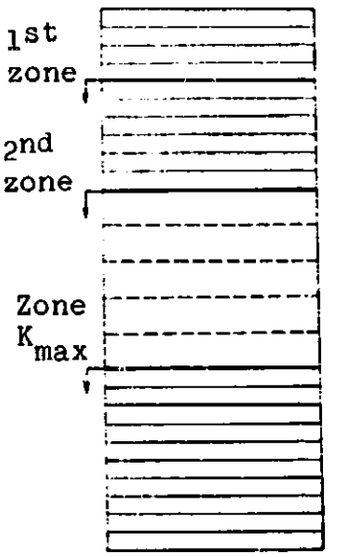


Figure 4. Table of ordered significant readings with mixed distribution

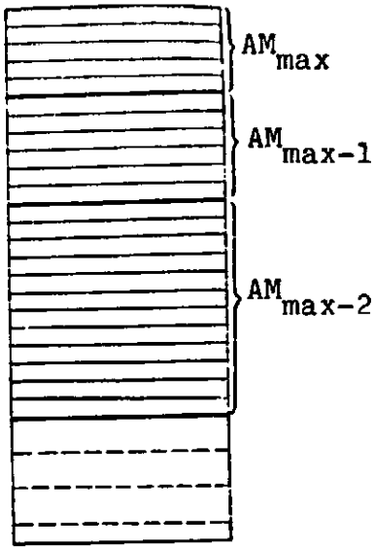


Figure 5. Table of ordered significant readings with floating distribution

transmitted significant reading table formation. In the general case, as mentioned previously, the principle of combining in each table of readings with a definite range of associative label values can be used as the basis of transmitted significant reading table formation.

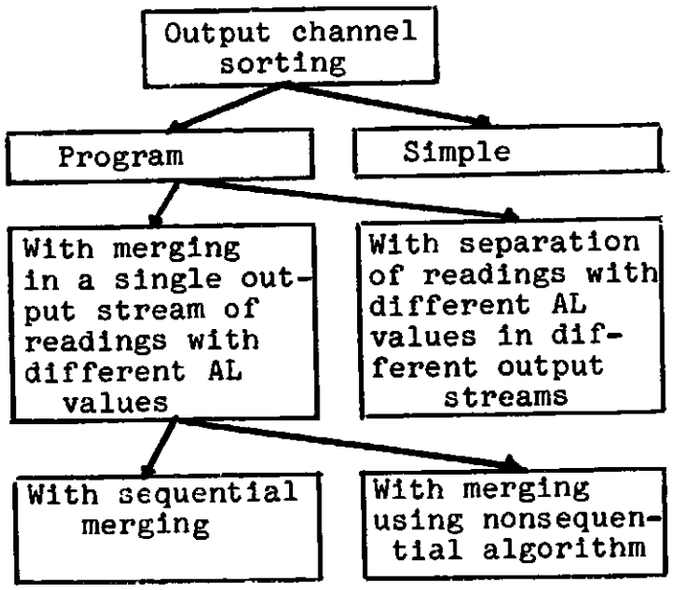


Figure 6. Classification of output channel sorting modes

Let us examine some variants of transmitted significant reading table formation:

1) each table includes significant readings with strictly fixed associative label value; the number of tables is equal to the number of possible associative label values (multiaddress organization of multipurpose information collection system);

2) a single table which includes readings with the complete range of associative label (multipriority organization of multipurpose systems);

3) R compressed message output streams or R transmitted significant reading tables. The first table includes readings with the largest associative label values in each formation cycle, the second includes those appearing in the first table and also readings with associative label values which are closest to the largest in the same formation cycle, and so on. The last table includes the maximal number of significant readings transmitted over the radio link during a time equal to the table formation cycle. The last table is intended for transmission over a channel with maximal information content. We call this type of output stream formation sorting with sequential merging;

4) R output streams, where formation of the reading tables for these streams takes place following a definite algorithm: in this case, we speak of algorithmic sorting.

Smoothers with complete and partial register sets (Figures 7, 8) can be used to smooth the nonuniformities of the compressed data output streams formed for transmission. In the first case, for readings with each associative label value, the matcher has as many output registers as there are output streams to which it is connected. In the second case, for readings with corresponding associative label value, there is a single output register; however, the control schemes are considerably more complex.

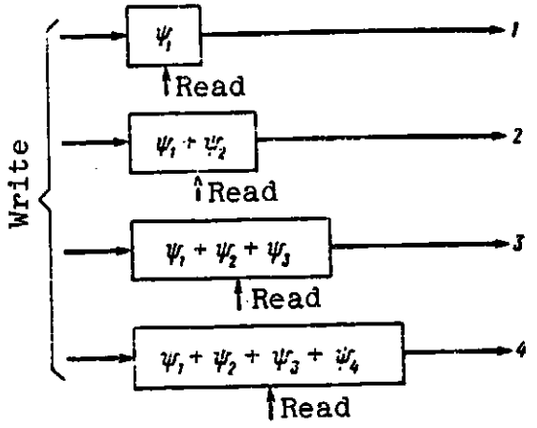


Figure 7. Matcher with complete register set

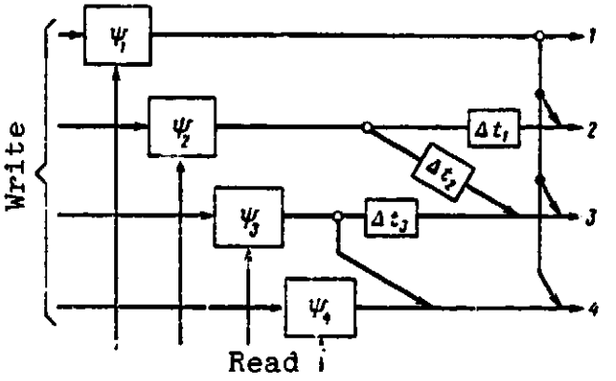


Figure 8. Matcher with incomplete register set

In conclusion, we note that preliminary calculations made by the present authors make it possible to hope that creation of multipurpose systems using AM will provide definite advantages from the viewpoint of size and weight characteristics in comparison with similar systems constructed by traditional methods.

REFERENCES

1. Kantor, A. V., V. G. Timonin and Yu. S. Azarova. In the collection: Apparatura dlya kosmicheskikh issledovaniy (Equipment for Space Studies). Nauka Press, Moscow, 1972.

2. Kantor, A. V., S. M. Perevertkin and T. S. Shcherbakova. In the collection: Apparatura dlya kosmicheskikh issledovaniy (Equipment for Space Studies). Nauka Press, Moscow, 1972.

ASSOCIATIVE COMPRESSED INFORMATION OUTPUT STREAM FORMATION BY THE STATISTICAL TRIAL METHOD

A. V. Kanton, T. A. Tolmadzheva

ABSTRACT. We examine a technique for studying the process of forming several compressed information streams, and present some results of such a study. We obtain the dependences of the output compressed information stream characteristics on the input stream characteristics, and also the dependences of the associative memory operation characteristics on the input and output stream characteristics.

References [1 - 2] considered the general questions of constructing multipurpose systems for the collection and processing of multichannel telemetry information and made an analytic study of output stream formation for a channel with maximal information content. Results were presented in [3] of modeling of multipurpose information collection systems with actual telemetry information which showed the advisability of more profound study of compressed information output stream formation. /26

Use of the statistical trial method [4] in studying multipurpose information collection systems makes it possible to model on a digital computer the process of output stream formation for a two-dimensional significant reading input stream with arbitrary distribution laws of both the time intervals between readings and the magnitudes of the approximation errors corresponding to each of these significant readings. Of significant importance for analysis of the output stream formation process is the establishment of the relationships between

the characteristics of the incoming significant reading stream and the characteristics of the output streams; at the same time, there is considerable interest in obtaining relations characterizing the operation of the associative memory, which is the basic element in realizing multipurpose information compression.

The present study is devoted to obtaining these relationships by the statistical trial method; we model on a digital computer one of the multipurpose compression modes (see [1]) — multichannel compression with number of output channels equal to the number of different approximation error values with floating significant reading distribution in the associative memory, and with constant duration of the multipurpose compression system operating cycle. The basic characteristics of the incoming sample stream are two random quantities: 1) z is the magnitude of the time interval between neighboring significant readings (for the simplest stream the equivalent quantity is the quantity B, characterizing the number of significant readings per multipurpose compression system operating cycle); 2) ε is the value of the approximation error, characterizing the significant sample. The two-dimensional incoming significant sample stream model is used in modeling the associative output stream formation processes [5]. The number B of significant samples per operating cycle is distributed in accordance with the Poisson law with parameter λ:

$$P_\tau(b) = \frac{(\lambda\tau)^b e^{-\lambda\tau}}{b!}, \qquad (1)$$

where b is the value of the random quantity B; τ is the operating cycle duration (τ = const); λτ is the average number of significant samples during the time τ; P_τ (b) is the probability that B = b. Therefore, the probability density P (z) of the time interval z between neighboring significant samples is

$$P(z) = \lambda e^{-\lambda z}. \qquad (2)$$

In the modeling, we assumed that the approximation error ε, characterizing each significant sample, is a discrete random quantity ε_K, given by the distribution series: /27

ε_1	ε_2	. . .	ε_K	. . .	$\varepsilon_{K\max}$,
P_1	P_2	. . .	P	. . .	$P_{K\max}$

where K_{max} is the maximal number of different error values. The studies were made for a distribution series of the form:

$$P_K = P_m e^{-CK}, \quad 1 \leqslant K \leqslant K_{max}. \tag{3}$$

With account for the normalization

$$\sum_{K=1}^{K_{max}} P_K = 1$$

Expression (3) is written as follows:

$$P_K = \frac{1}{\sum_{K=1}^{K_{max}} e^{-CK}} e^{-CK}. \tag{4}$$

In the modeling process, the incoming stream is represented in the form of the sum of K_{max} one-dimensional streams. The characteristics of each of these streams are the quantities n_{ε_K}, defining the number of significant readings with approximation error $\varepsilon = \varepsilon_K$, written in the associative memory during the r^{th} operating cycle, i.e., during the time τ. Then,

$$\lambda_K \tau = M\{n_{\varepsilon_K}\} = \frac{\sum_{r=1}^{r_{max}} n_{\varepsilon_K}}{r_{max}}, \tag{5}$$

where λ_K is the average density of the significant sample stream with error ε_K. It is obvious that

$$\lambda = \sum_{K=1}^{K_{max}} \lambda_K. \tag{6}$$

The basic characteristics of each of the R output streams are:

1) $(f)_{R-\beta}$ — the clock frequency, equal to the significant sample repetition rate in the absence of dropouts; $(R - \beta)$ — the stream number, $0 \leq \beta \leq R - 1$;

2) $(\varepsilon)_{R-\beta}$ — approximation error, characterized by the following random quantities:

a) $(\varepsilon_{nb})_{R-\beta}$ — approximation pseudo-error, values of the error ε of samples entering the output stream with number $R - \beta$;

b) $(\varepsilon_{+1})_{R-\beta}$ — maximal value of the error ε of samples entering the output stream with number $R - \beta + 1$;

c) $(\varepsilon_{n})_{R-\beta}$ — value of the error ε of samples which are not transferred to the output stream with number $R - \beta$;

3) $(K_{con})_{R-\beta}$ — contraction coefficient of the clock frequency in the stream with number $R - \beta$ relative to the stream with number R. The quantity $(f)_{R-\beta}$ is defined by the relation

$$(f)_{R-\beta} = \frac{\sum_{\gamma=1}^{R-\beta} \varphi_\gamma}{\Pi} \frac{1}{\tau}, \quad 0 \leqslant \beta \leqslant R-1, \tag{7}$$

where $\phi_{R-\beta}$ is the number of associative memory cells from which information is added to the information of the stream $R - \beta - 1$ to obtain the stream $R - \beta$; /28

$$\sum_{\gamma=1}^{R} \varphi_\gamma = \Pi$$

is the maximal number of significant samples entering the R^{th} output stream during each cycle (see below);

$$(K_{con})_{R-\beta} = \frac{\Pi}{\sum_{\gamma=1}^{R-\beta} \varphi_\gamma} = \frac{1}{(f)_{R-\beta}\tau}. \tag{8}$$

It is obvious that for the R^{th} stream, $(K_{con})_R = 1$.

Additional characteristics of the R^{th} stream will be the quantities S, defining the number of ordered significant samples actually transferred to the R^{th} channel in the r^{th} operating cycle, i.e., during the time τ.

The basic characteristics and parameters of AM operation are:

1) the characteristics of the random quantity $M_{\varepsilon K}$, defining the number of significant readings stored in the associative memory immediately prior to initiation of sampling in the r^{th} operating cycle; it is obvious that

$$m_{\varepsilon_K} = n_{\varepsilon_K} + m'_{\varepsilon_K},$$

where $m'_{\varepsilon K}$ is the number of significant readings which are not transferred in the $(r - 1)^{th}$ operating cycle;

2) the characteristics of the random quantity m_ε, characterizing the length of the reading sequence, where

$$m_\varepsilon = \sum_{K=1}^{K_{max}} m_{\varepsilon_K};$$

3) τ — operating cycle duration;

4) f_R — clock frequency in the R^{th} output stream;

5) Π — maximal number of significant samples entering the R^{th} output stream during each cycle; it is obvious that

$$S = \begin{cases} \Pi \text{ for } m_\varepsilon \geqslant \Pi, \ f_R = \frac{\Pi}{\tau}; \\ m_\varepsilon \text{ for } m_\varepsilon < \Pi, \end{cases} \tag{9}$$

6) P_{ov} — probability of AM overflow:

$$P_{ov} = P\{m_\varepsilon > E_{AM}\}, \tag{10}$$

where E_{AM} is the AM capacity;

7) $(P_{ov})_{R-\beta}$ — probability of overflow of the associative memory zone, the information in the cells of which is used to form

the stream with number R - β:

$$(P_{em})_{R-\beta} = P\left\{\left(\sum_{\gamma=\beta+1}^{K_{max}} m_{\varepsilon_\gamma}\right) > \left(\sum_{\gamma=1}^{R-\beta} \varphi_\gamma\right)\right\}; \quad (11)$$

8) P_{em} — probability of AM emptying:

$$P_{ov} = P\{S < \Pi\} = 1 - P\{S > \Pi\}. \quad (12)$$

The sought relations are shown in the table. /29

TABLE

Argument	Function	Parameters	Constants
$K_{con\ 1}$	Characteristics of random quantities n_{ε_K}; m_{ε_K}; ε_{nb}; ε_{+1}; ε_n; S	C/C_n	ρ
$K_{con\ 1}$	Quantities: P_{ov}; P_{em}; $(P_{ov})_{R-\beta}$	ρ	C/C_n
ρ		Π	C/C_n
C/C_n		$K_{con\ 1}$	ρ
ρ		C/C_n	$K_{con\ 1}$

The following notations are used in the table: C_n is the nominal value of the coefficient C in (4); the nominal value of C is defined by the relation /32

$$K_{con\ 1} = \frac{1}{P_{K\,max}} = \frac{\sum_{K=1}^{K_{max}} e^{-KC}}{e^{-K_{max}C_n}}. \quad (13)$$

This expression is obtained from (4) and (8) for β = R - 1, and ρ = 1; ρ is the load, defined by the expression

$$\rho = \frac{\lambda\tau}{\Pi}. \quad (14)$$

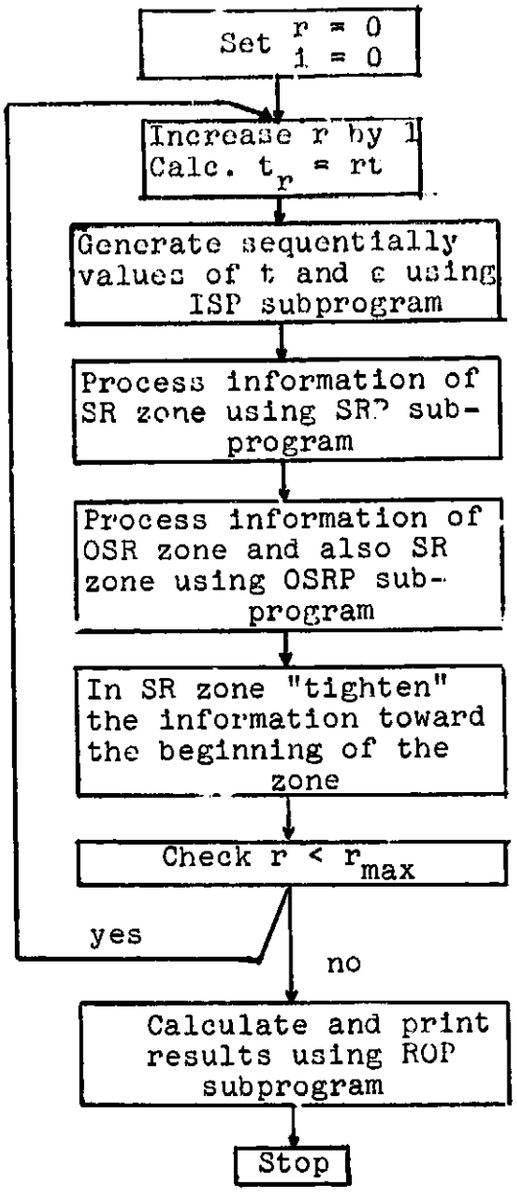


Figure 1. Block diagram of algorithm of program for studying output stream formation process

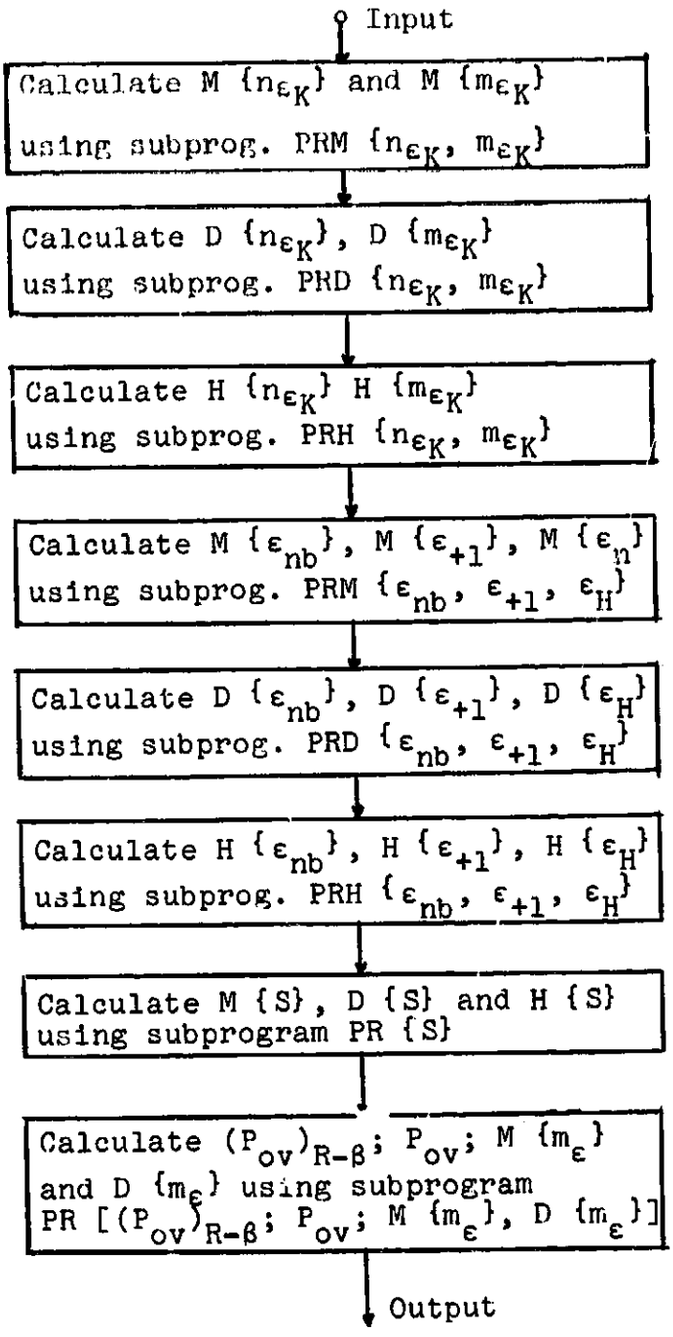


Figure 2. Block diagram of algorithm of input stream formation subprogram

We shall describe the block diagram of the algorithm for studying the output stream formation process. Figure 1 shows the expanded block diagram of the program algorithm for studying the output stream formation process. The program includes the following basic subprograms: the input stream formation program (ISP); the significant reading zone information processing program (SRP); the ordered significant reading zone information processing program (OSRP); and the result output program (ROP). The following notations are used in Figure 1 [in addition to those indicated in the text and the Relations (1) - (12) presented above]: i is the number of the random number in the ISP subprogram (Figure 2); t_r is the r^{th} cycle termination time; r_{max} is the number of cycles.

In each operating cycle r the significant readings are generated using the ISP subprogram (Figure 2), the time intervals z_i are determined by (2), and the values of the approximation error ε_K are determined by (4). The quantities z_i and ε_K are formed from the random numbers x_i with uniform probability distribution in the interval 0 - 1, generated by a digital computer using a standard program. Upon termination of the interval $t_{r-1} - t_r$, stream formation ends. However, the possibility of appearance of the last reading after the moment t_r is not excluded. This reading is stored in the transient reading (TR) cell, and is used in the next cycle. The values of the reading occurrence time $t = t_i - t_{r-1}$ (t_i is the time of reading occurrence after the interval z_i) and the error magnitudes ε_K are stored in the machine memory in the order of their entry into the zone, called the significant reading entry zone. Then the information is processed using the basic SRP subprogram (Figure 3) and the auxiliary subprograms in the SRP for calculating the mathematical expectation M $\{n_{\varepsilon K}\}$ and M $\{m_{\varepsilon K}\}$ (Figure 4), histogram H $\{n_{\varepsilon K}\}$ and H $\{m_{\varepsilon K}\}$, dispersion D $\{n_{\varepsilon K}\}$ and D $\{m_{\varepsilon K}\}$, and also the probability (11) of overflow of the zone (R - β), $(P_{ov})_{R-\beta}$, mathematical

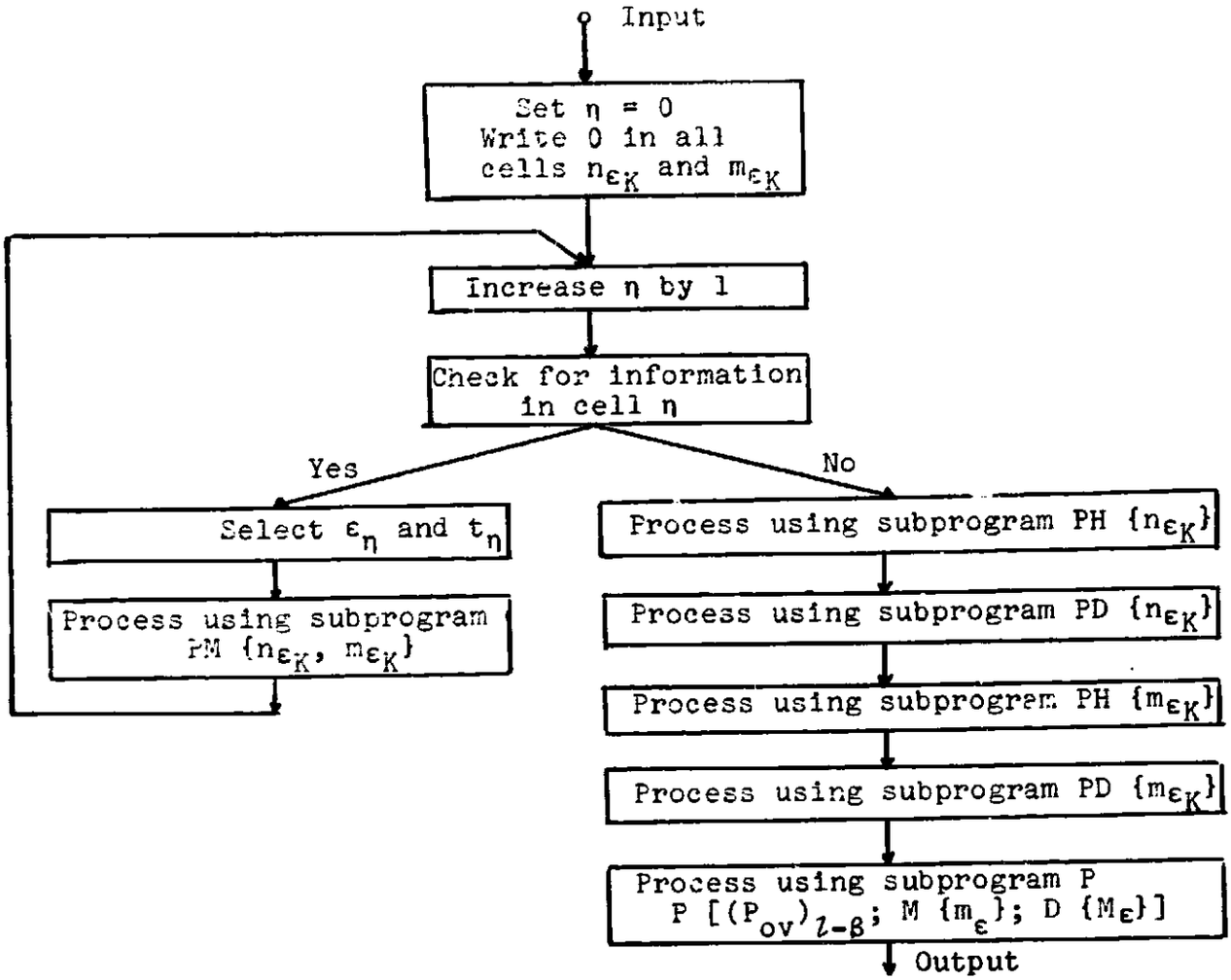


Figure 3. Block diagram of subprogram algorithm for processing information in SR zone

expectation M $\{m_\varepsilon\}$ and dispersion D $\{m_\varepsilon\}$ (Figure 5). The processing proceeds sequentially by rows (cells) η. Upon termination of processing, formation of the ordered significant readings (OSR) zone is accomplished, i.e., entry into the individual digital computer memory zone of the information from the significant reading zone (without destruction): the information is arranged in the ordered significant reading zone in order of decreasing the ε values, and for equal ε in order of increasing t. This process simulates associative information sorting in the AM. Then information

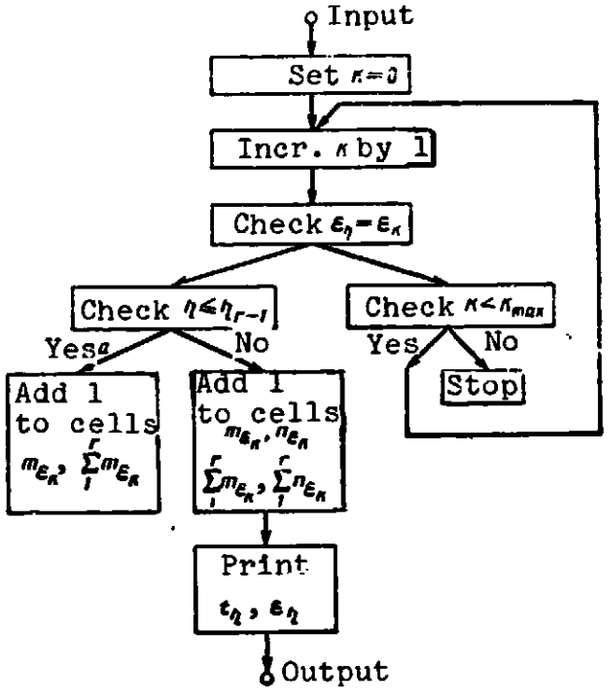


Figure 4. Block diagram of algorithm of subprogram for calculating M $\{n_{\varepsilon_K}\}$ and M $\{m_{\varepsilon_K}\}$

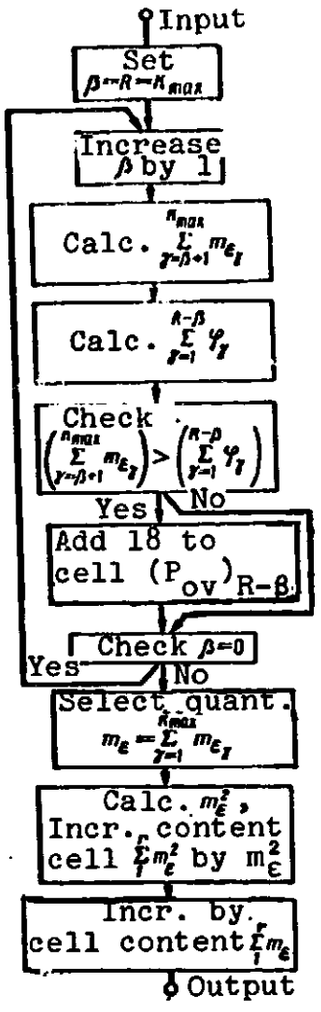


Figure 5. Block diagram of algorithm of subprogram for calculating $(P_{ov})_{R-\beta}$, M $\{m_\varepsilon\}$ and D $\{m_\varepsilon\}$

processing is accomplished using the basic OSRP subprogram (Figure 6), and several auxiliary subprograms. The information is processed sequentially by rows (cells) q and channels β. This subprogram is used to calculate the characteristics of the random quantities ε_{nb}; ε_n; ε_{+1}; S, and the probability of AM capacity overflow P_{ov} [see (10)], and emptying P_{em} [see (12)].

Upon termination of the program in the significant reading zone, the remaining information is "tightened" toward the beginning of the significant sample entry zone, i.e., the cells which have been

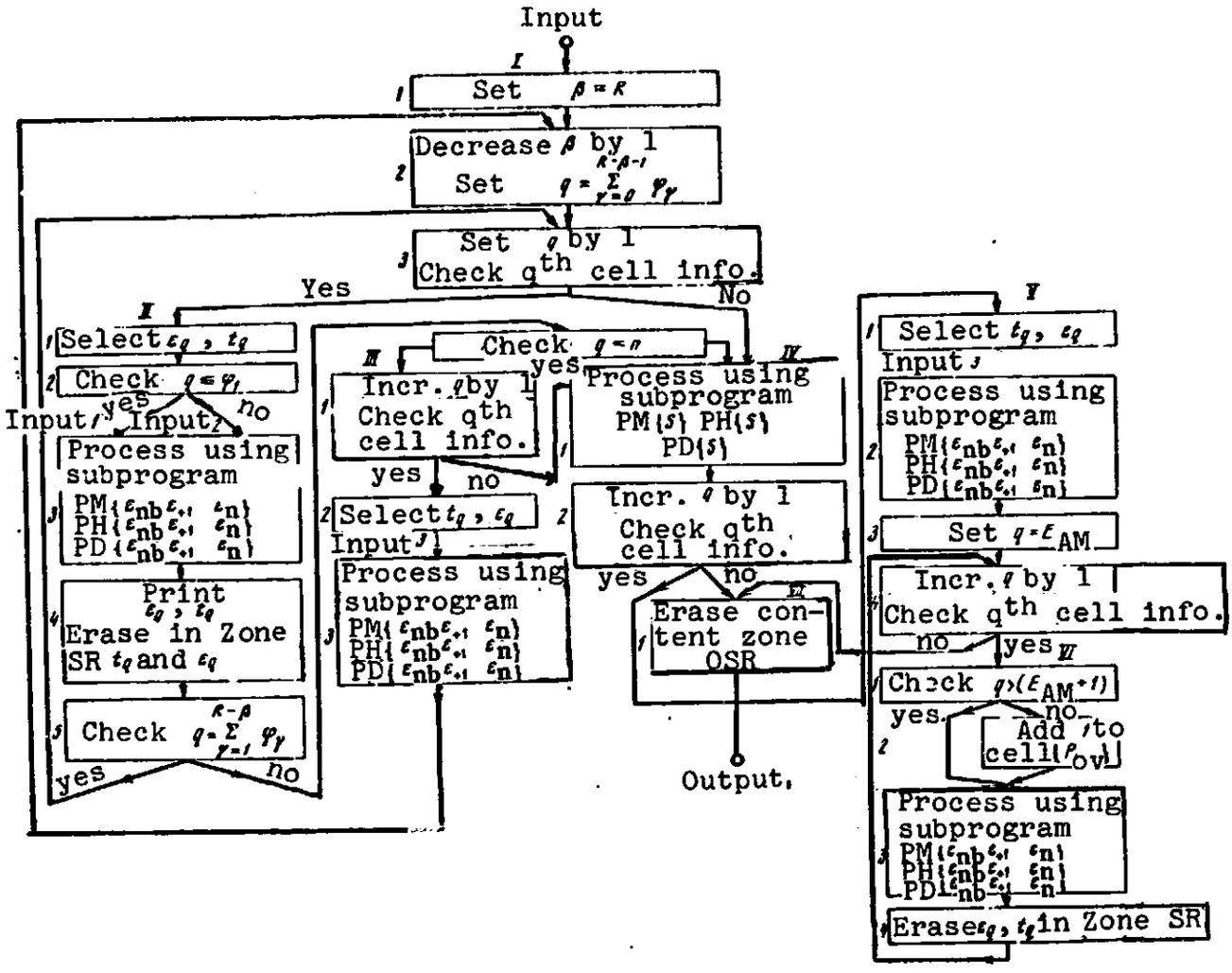


Figure 6. Block diagram of algorithm of subprogram for processing information in OSR zone

vacated are filled. This terminates modeling of the r^{th} operating cycle, and transfer to the $(r + 1)^{th}$ cycle takes place. Upon termination of information processing using the subprograms ISP, SRP, OSRP ($r = r_{max}$), the results are processed using the basic ROP subprogram (Figure 7) and several auxiliary subprograms, specifically, the subprogram for calculating $(P_{ov})_{R-\beta}$ and P_{ov} (Figure 8). During processing using the ISP, SRP, OSRP subprograms, data are

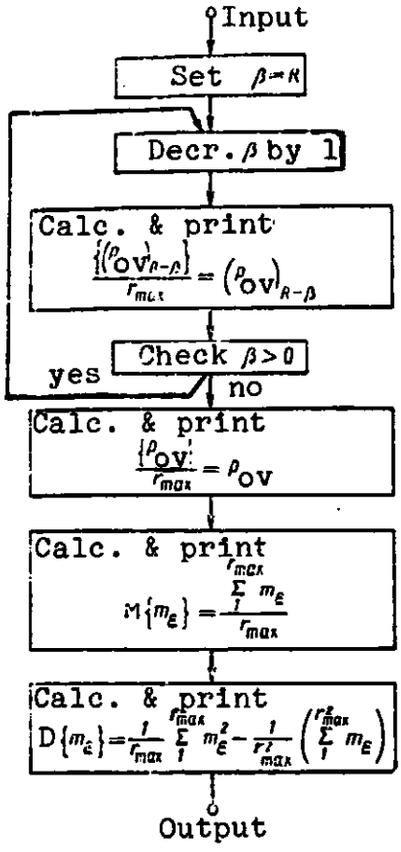


Figure 7. Block diagram of algorithm subprogram for printout result

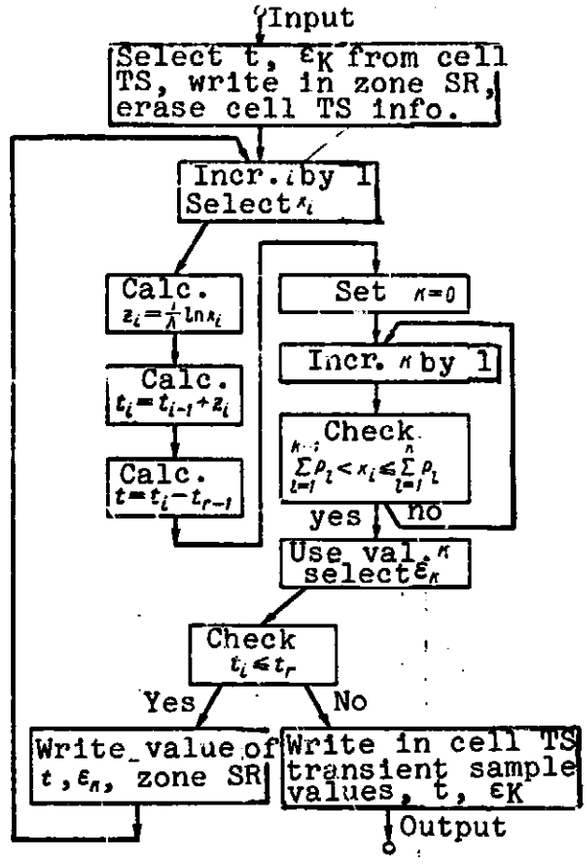


Figure 8. Block diagram of algorithm of subprogram for calculating $(P_{ov})_{R-\beta}$ and P_{ov}

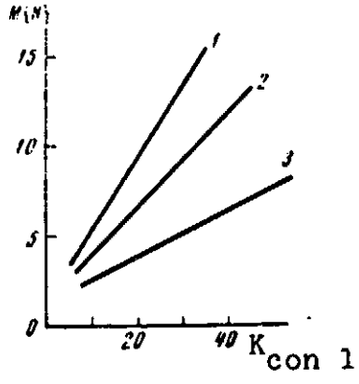


Figure 9. Formation process characteristics versus contraction factor

1 — $C/C_n = 0.8$; $\rho = 1$; 2 — $C/C_n = 1$; $\rho = 0.95$; 3 — $C/C_n = 1, 2$; $\rho = 0.7$

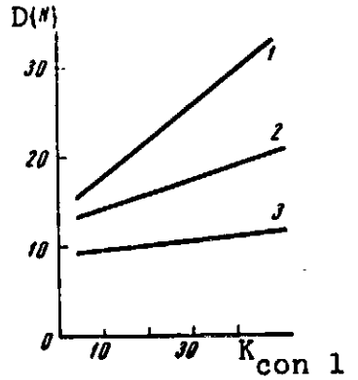


Figure 10. Formation process characteristics versus contraction factor

1 — $\rho = 1$; $C/C_n = 0.8$; 2 — $\rho = 0.95$; $C/C_n = 1$; 3 — $\rho = 0.7$; $C/C_n = 1, 2$

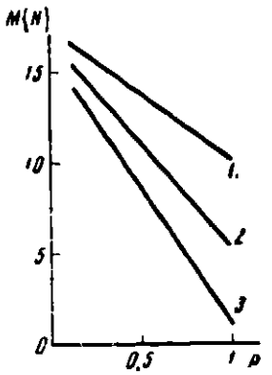


Figure 11. Formation process characteristics versus loading coefficient ρ

1 — Π = 16; C/C_n = 0.8; 2 — Π = 32; C/C_n = 1; 3 — Π = 48; C/C_n = 1, 2

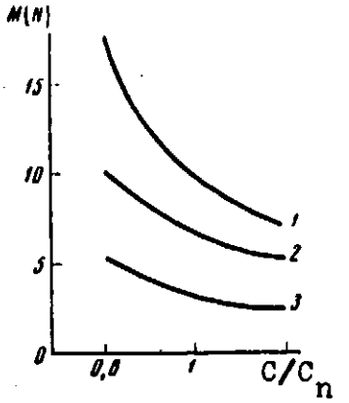


Figure 12. Formation process characteristics versus ratio C/C_n

1 — $K_{con\ 1}$ = 8; ρ = 1; 2 — $K_{con\ 1}$ = 16; ρ = 0.95; 3 — $K_{con\ 1}$ = 32; ρ = 0.7

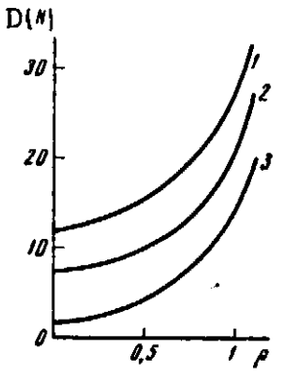


Figure 13. Formation process characteristics versus loading coefficient ρ

1 — C/C_n = 0.8; $K_{con\ 1}$ = 8; 2 — C/C_n = 1; $K_{con\ 1}$ = 16; 3 — C/C_n = 1, 2; $K_{con\ 1}$ = 32

accumulated for calculating the mathematical expectation, dispersion, and histogram of the random quantities n_{ε_K}; m_{ε_K}; ε_{nb}; ε_n; ε_{+1}; S, and data for $(P_{ov})_{R-\beta}$ and P_{ov}. The ROP subprogram is used to calculate the aforementioned characteristics based on the operating cycle r_{max} values.

The results of the digital computer study of associative output stream formation are shown in Figures 9 - 13. The data obtained reflect quite completely the relationships between the incoming significant reading stream characteristics and the output stream characteristics, and also characterize the associative memory operation.

In conclusion, the authors wish to thank I. I. Golovchenko and R. A. Sagitov for their assistance in developing the program and making the computer calculations.

REFERENCES

1. Kantor, A. V., S. M. Perevertkin and T. S. Shcherbakova. See present volume, p. 34.

2. Kantor, A. V., S. M. Perevertkin and T. S. Shcherbakova. See present volume, p. 57.

3. Tolmadzheva, T. A., A. V. Kantor and L. V. Rozhkovskiy. In the collection: Apparatura dyla kosmicheskikh issledovaniy (Equipment for Space Studies). Nauka Press, Moscow, 1972.

4. Buslenko, N. P. Modelirovaniye slozhnykh sistem (Modeling Complex Systems). Nauka Press, Moscow, 1968.

5. Kantor, A. V. and T. A. Tolmadzheva. In the collection: Apparatura dlya kosmicheskikh issledovaniy (Equipment for Space Studies). Nauka Press, Moscow, 1972.

ANALYTIC STUDY OF OUTPUT STREAM FORMATION PROCESS IN MULTIPURPOSE INFORMATION COMPRESSION SYSTEMS

A. V. Kantor, S. M. Perevertkin and
T. S. Shcherbakova

ABSTRACT. An analytic study is made of the output stream formation process for a system with maximal information content. We take as the mathematical model the single-channel mass servicing system with very simple input stream, with group demand servicing, and Erlangian servicing interval distribution. Analytic expressions are obtained for several system state probability distributions and queue length mathematical expectation and dispersion which can be used for certain simplified calculations of the formation process of all the output streams, but with the use of priority servicing and with a simplified servicing interval model.

As was shown in [1], it is advisable to study the multipurpose /35
information compression system by queueing theory methods. In this case, such a system can be considered a queueing system with waiting. A stream of significant readings with corresponding associative labels enters the input of the associative memory (AM) [2] used as a sorter and buffer memory (BM) in multipurpose information compression systems. The significant reading stream is taken to be very simple (experimental studies confirm the validity of this assumption for stationary segments of telemetry parameter behavior.

The output of readings from the AM, i.e., servicing, is accomplished by groups of П readings during the multipurpose information compression system operating cycle.

In this study, the significant reading group servicing time is assumed to be distributed in accordance with the Erlang law with density

$$f(\tau) = \frac{\mu^K \tau^{K-1}}{\Gamma(K)} e^{-\mu\tau}, \qquad (1)$$

where $\frac{\mu}{K} = \frac{1}{M\{\tau\}}$ is the average number of significant readings taken from the AM into the matcher register per unit time; K is the parameter of the Erlang distribution. Specifically, for $K = 1$ $f(\tau) = \mu e^{-\mu\tau}$, which corresponds to the exponential distribution, for K = ∞ we have the constant servicing duration case examined in [1].

In analyzing the states of the subject mass servicing system, we use the embedded Markov chain method [3] in which the system states are examined at strictly defined moments of time, namely, immediately preceding the moments of servicing termination. We note that here, as usual, by system states we mean the number of messages (demands) in the associative memory at a definite moment of time.

The Chapman-Kolmogorov equations are valid for embedded Markov chain points:

$$P_n(j) = \sum_l P_l(j-1)\, P_{ln}(j), \qquad (2)$$

where $P_n(j)$ is the probability that at the end of the j^{th} cycle, the system will be in tne state n; $P_l(j-1)$ is the probability that at the end of the (j - 1)th cycle, the system will be in the state l; $P_{ln}(j)$ is the probability of process transition from state l into state n, where

$$P_{ln}(j) = P\{N(j) = n / N(j-1) = l\}.$$

Let q_b be the probability that in the j^{th} cycle, the number of newly arriving readings will be b. It is obvious that for $0 \leqslant l \leqslant H$, b = n, for $H < l \leqslant H + n$, $b = n - l + H$, and for $l > H + n$, b does not define the system state at the end of the j^{th} cycle, i.e.,

$$P_{ln}\begin{cases} q_n & \text{for } 0 \leqslant l \leqslant H-1, \\ q_{n-l}+H & \text{for } H \leqslant l \leqslant H+n, \\ 0 & \text{for } H+n < l < \infty. \end{cases} \tag{3}$$

The corresponding transition probability matrix has the form (the subscript j is omitted for simplification): /36

$$\|P_{ln}\| = \begin{vmatrix} q_0 & q_1 & q_2 & q_3 \cdots \\ q_0 & q_1 & q_2 & q_3 \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ q_0 & q_1 & q_2 & q_3 \cdots \\ q_0 & q_1 & q_2 & q_3 \cdots \\ 0 & q_0 & q_1 & q_2 \cdots \\ 0 & 0 & q_0 & q_1 \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 \ldots \end{vmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \} H-1 \\ \\ \\ \\ \} n+1 \\ \\ \\ \\ \\ \end{matrix}$$

With account for (3), Equation (2) is written as follows:

$$\begin{aligned} P_n(j) &= \sum_{l=0}^{n-1} P_l(j-1)\,q_n(j) + \sum_{l=H}^{H+n} P_l(j-1)\,q_{n-l+n}(j) = \\ &= q_n(j) \sum_{l=0}^{n-1} P_l(j-1) + \sum_{l=H}^{H+n} P_l(j-1)\,q_{n-l+H}(j), \\ \infty &\geqslant n \geqslant 0. \end{aligned} \tag{4}$$

The probabilities P_n (j) satisfy the normalization condition

$$\sum_{n=0}^{\infty} P_n(j) = 1. \tag{5}$$

For the steady-state regime $(j \to \infty)$, the system of equations (4) will have the form

$$\begin{aligned} P_0 &= q_0 \sum_{l=0}^{H} P_l + P_H q_0, \\ P_1 &= q_1 \sum_{l=0}^{H-1} P_l + P_H q_1 + P_{H+1} q_0, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ P_H &= q_n \sum_{l=0}^{H-1} P_l + \sum_{l=0}^{H+n} P_l q_{n-l+H}, \qquad \sum_{n=0}^{\infty} P_n = 1. \end{aligned} \tag{6}$$

In solving the system of equations (6), we use the generating function method. We introduce the generating functions:

$$P(Z) = \sum_{n=0}^{\infty} P_n Z^n, \tag{7}$$

$$Q(Z) = \sum_{n=0}^{\infty} q_n Z^n. \tag{8}$$

Substituting the values of P_n from (6) into (7), we obtain

$$\begin{aligned} P(Z) &= \sum_{n=0}^{\infty} Z^n \Big[q_n \sum_{l=0}^{H-1} P_l + \sum_{l=H}^{H+n} P_l q_{n-l+H} \Big] = \\ &= \sum_{n=0}^{\infty} q_n Z^n \sum_{l=0}^{n-1} P_l + \sum_{n=0}^{\infty} Z^n (P_H q_n + P_{H+1} q_{n-1} \dots + P_{H+n} q_0) = \\ &= Q(Z) \sum_{l=0}^{H-1} P_l + \sum_{n=0}^{\infty} Z^n (P_H q_n + P_{H+1} q_{n-1} + \dots + P_{H+n} q_0). \end{aligned} \tag{9}$$

The convolution of the two sequences $\{q_n\}$ and $\{P_{H+n}\}$ appears in the brackets in the right side of (9). Unwrapping this convolution, we obtain: /37

$$P(Z) = Q(Z) \sum_{l=0}^{H-1} P_l + \Big(\sum_{n=0}^{\infty} Z^n q_n \Big) \sum_{n=0}^{\infty} Z^n P_{H+n}, \tag{10}$$

where

$$\sum_{n=0}^{\infty} Z^n q_n = Q(Z). \tag{11}$$

The sum $\sum_{n=0}^{\infty} Z^n P_{H+n}$ in (10) can be transformed as follows:

$$\sum_{n=0}^{\infty} Z^n P_{H+n} = Z^{-H} \Big[P(Z) - \sum_{l=0}^{H-1} P_l Z^l \Big]. \tag{12}$$

The Expression (9) can be written in the form

$$P(Z) = Q(Z) \sum_{l=0}^{H-1} P_l + Q(Z) Z^{-H} \Big[P(Z) - \sum_{l=0}^{H-1} P_l Z^l \Big]. \tag{13}$$

Solving (13) for P (Z), we obtain:

$$P(Z) = \frac{\sum_{l=0}^{H-1} P_l (Z^H - Z^l)}{\frac{Z^H}{Q(Z)} - 1}. \tag{14}$$

In the Erlangian servicing time distribution case, the Expression (8) for Q (Z) takes the form:

$$Q(Z) = \left(1 + \frac{\lambda(1-Z)}{\mu}\right)^{-K} = \left(1 + \frac{\rho\Pi(1-Z)}{K}\right)^{-K}, \qquad (15)$$

where $\rho = \lambda K / \mu\Pi$ is the load; K is the parameter of the Erlangian distribution; λ is the incoming stream intensity; Π is the number of significant readings in the group. Then (14) takes the form:

$$P(Z) = \frac{\sum_{i=0}^{\Pi-1} P_i (Z^{\Pi} Z^i)}{Z^{\Pi}\left(1 + \frac{\rho\Pi(1-Z)}{K}\right)^{K} - 1}. \qquad (16)$$

Omitting the intermediate transformations, (16) may be written in the form:

$$P(Z) = \frac{C}{\prod_{j=n}^{\Pi+K-1} (Z_j - Z)}, \qquad (17)$$

where C is a constant; Z_j are the roots of the denominator of (16) outside the unit circle. The constant C is determined from the condition P (Z) = 1 for Z = 1, hence:

$$C = \prod_{j=n}^{\Pi+K-1} (Z_j - 1) \qquad (18)$$

Finally, we write the expression for the generating function P (Z): /38

$$P(Z) = \prod_{j=\Pi}^{\Pi+K-1} \frac{Z_j - 1}{Z_j - Z}. \qquad (19)$$

To find the system state probabilities, we expand the expression for P (Z) into simple fractions:

$$P(Z) = \frac{\beta_{\Pi}}{Z_{\Pi} - Z} + \frac{\beta_{\Pi+1}}{Z_{\Pi+1} - Z} + \cdots \frac{\beta_{\nu}}{Z_{\nu} - Z} + \cdots \frac{\beta_{\Pi+K-1}}{Z_{\Pi+K-1} - Z}, \qquad (20)$$

where the expansion coefficients β_{ν} are found from the expression:

$$\beta_{\nu} = \frac{-\prod_{j=\Pi}^{\Pi+K-1} (Z_j - 1)}{\frac{d}{dZ} \prod_{j=\Pi}^{\Pi+K-1} (Z_j - Z) \Big|_{Z=Z_{\nu}}} \qquad (21)$$

Varying ν in the limits $H + K - 1 \geqslant \nu \geqslant H$, and substituting the corresponding expressions for (21) into (20), we have:

$$P(Z) = \frac{\beta_H}{Z_H}\left[1 + \frac{Z}{Z_H} + \left(\frac{Z}{Z_H}\right)^2 + \left(\frac{Z}{Z_H}\right)^3 + \dots\right] + \frac{\beta_{H+1}}{Z_{H+1}}\left[1 + \frac{Z}{Z_{H+1}} + \left(\frac{Z}{Z_{H+1}}\right)^2 + \left(\frac{Z}{Z_{H+1}}\right)^3 + \dots\right] + \frac{\beta_{H+K-1}}{Z_{H+K-1}}\left[1 + \frac{Z}{Z_{H+K-1}} + \left(\frac{Z}{Z_{H+K-1}}\right) + \dots\right]. \quad (22)$$

We obtain the expression for P_n from (21) as coefficients of the corresponding powers of Z. Thus,

$$P_n = \frac{\beta_H}{Z_H^{n+1}} + \frac{\beta_{H+1}}{Z_{H+1}^{n+1}} + \dots + \frac{\beta_{H+K-1}}{Z_{H+K-1}^{n+1}}. \quad (23)$$

If Z_H is the smallest-in-magnitude root of the denominator of (16) outside the unit circle, then the approximate expression for P_n has the form:

$$P_n \simeq \frac{\beta_H}{Z_H^{n+1}}. \quad (24)$$

The mathematical expectation of the number of demands in the system and the dispersion of the number of demands are found from (22):

$$M\{N\} = \frac{dP(Z)}{dZ}\Big|_{Z=1} \quad (25)$$

and

$$D\{N\} = \frac{d^2P(Z)}{dZ^2}\Big|_{Z=1} + \frac{d^2P(Z)}{dZ^2}\Big|_{Z=1} - \left(\frac{dP(Z)}{dZ}\Big|_{Z=1}\right). \quad (26)$$

Direct calculation of the first and second moments using (25) and (26) is difficult. We use the semi-invariant method of [4]. Setting $Z = l^\theta$, we have

$$M(\theta) = \sum_{n=0}^{\infty} P_n l^{\theta n}, \quad (27)$$

where M (θ) is the moment generating function.

Taking the logarithm of (27), we obtain the generating function /39 of the semi-invariants ψ (θ). The mathematical expectation of the number of significant readings in the system and their dispersion are obtained by differentiating the semi-invariant generating

function, setting $\theta = 0$:

$$M\{N\} = \frac{d\psi(0)}{d\theta}\Big|_{\theta=0} = \sum_{j=H}^{H+K-1} (Z_j - 1)^{-1}, \tag{28}$$

$$D\{N\} = \frac{d^2\psi(0)}{d\theta^2}\Big|_{\theta=0} = \sum_{j=H}^{H+K-1} Z_j (Z_j - 1)^{-2}. \tag{29}$$

Let us examine some limiting Erlang distributions:

1) $K = 1, 1 < H < \infty$ (servicing interval distributed exponentially);

2) $K = \infty, 1 < H < \infty$ (servicing interval constant);

3) $H = 1, 1 \leqslant K \leqslant \infty$ (single servicing);

4) $H = \infty, 1 \leqslant K \leqslant \infty$ (group servicing with servicing after a single interval of an infinitely large number of SR).

For the limiting Erlang distribution cases presented above, we obtain the following expressions for the generating function, distribution series, mathematical expectation, queue length, and queue length dispersion:

$$1.\ P(Z) = \frac{Z_H - 1}{Z_H - Z}, \tag{30}$$

$$P_n = \frac{Z_H - 1}{Z_H^{n+1}}. \tag{31}$$

Here, Z_H is the single zero outside the unit circle of the denominator of (16):

$$M\{N\} = \frac{1}{Z_H - 1}, \tag{32}$$

$$D\{N\} = \frac{Z_H}{(Z_H - 1)^2}. \tag{33}$$

$$2.\ P(Z) = \frac{H(1-\rho)(Z-1)\prod_{i=1}^{H-1}\frac{Z - Z_i}{(1 - Z_i)}}{Z_i^H e^{\rho H(1-Z)}}, \tag{34}$$

where Z_i are the roots of the denominator inside the unit circle and on its boundary.

$$p_n = (-1)^{H-n+1} a_{H-n} \frac{H(1-\rho)}{\prod_{i=1}^{H-1}(1-Z_i)}, \quad n \leqslant H-1, \tag{35}$$

$$a_{H-n} = \sum_{i<K<l<H} \frac{Z_i Z_K \dots Z_H}{(H-n)},$$

$$M\{N\} = \frac{1 - H(1-\rho)^2}{2(1-\rho)} + \sum_{i=1}^{H-1} \frac{1}{(1-Z_i)}, \tag{36}$$

$$D\{N\} = \frac{(1+2\rho) + 6\rho H(1-\rho)^2 - H^2(1-\rho)^2}{12(1-\rho)} - \sum_{i=1}^{H-1} \frac{Z_i}{(1-Z_i)^2}. \tag{37}$$

3. a) $1 < K < \infty$.

$$P(Z) = \prod_{j=1}^{K} \frac{Z_j - 1}{Z_j - Z}, \tag{38}$$ /40

$$p_n = \sum_{\nu=1}^{K} \frac{\beta_\nu}{Z_\nu^{n+1}}, \quad \beta_\nu = \frac{\prod_{j=1}^{K}(Z_j - 1)}{\frac{d}{dZ}\prod_{j=1}^{K}(Z_j - Z)\Big|_{Z=Z_\nu}}, \tag{39}$$

$$M\{N\} = \sum_{j=1}^{K} (Z_j - 1)^{-1}, \tag{40}$$

$$D\{N\} = \sum_{j=1}^{K} Z_j (Z_j - 1)^{-2}. \tag{41}$$

b) $H = 1, \quad K = 1.$

$$P(Z) = \frac{Z_1 - 1}{Z_1 - Z}, \tag{42}$$

$$p_n = \frac{\beta_1}{Z_1^{n+1}} = \frac{Z_1 - 1}{Z_1^{n+1}}, \tag{43}$$

$$M\{N\} = \frac{1}{Z_1 - 1}, \tag{44}$$

$$D\{N\} = \frac{Z_1}{(Z_1 - 1)^2}. \tag{45}$$

c) $H = 1, \quad K = \infty,$

$$P(Z) = \frac{(1-\rho)(Z-1)}{Ze^{\rho(1-Z)} - 1}, \tag{46}$$

$$p_n = \begin{cases} 1-\rho & \text{for } n = 0, \\ (1-\rho)(e^{\rho} - 1) & \text{for } n = 1, \\ (1-\rho)\sum_{S=1}^{n} (-1)^{n-S} e^{S\rho} \left[\frac{(S\rho)^{n-S}}{(n-S)!} + \frac{(S\rho)^{n-S-1}}{(n-S-1)!} \right] & \text{for } n \geqslant 2, \end{cases} \tag{47}$$

$$M\{N\} = \frac{\rho(2-\rho)}{2(1-\rho)}, \tag{48}$$

$$D\{N\} = \frac{(1+2\rho) + 6\rho(1-\rho)^2 - (1-\rho)^4}{12(1-\rho)}. \tag{49}$$

4. a) $H = \infty, \quad 1 < K < \infty$.

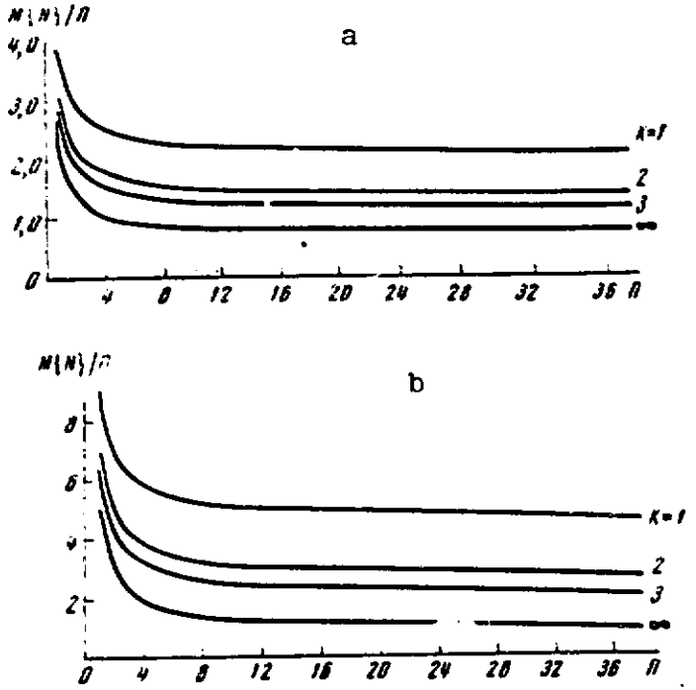


Figure 1. Average queue length versus number of readings in group:

a — ρ = 0.8; b — ρ = 0.9

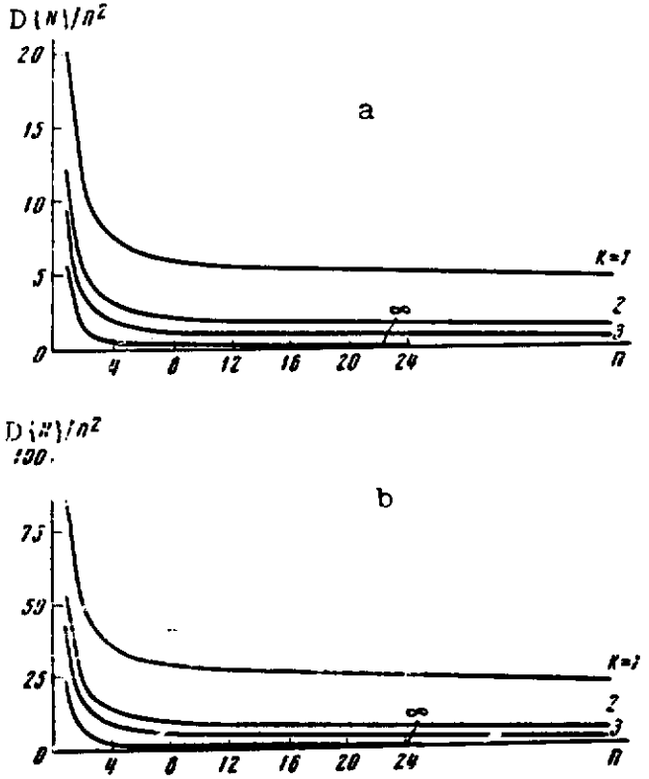


Figure 2. Queue length dispersion versus number of readings in group:

a — ρ = 0.8; b — ρ = 0.9

Upon substituting into (19), in place of Z, the quantity (1 - Z/Π), we obtain the expression for the generating function P (Z) [5]:

$$P_{\Pi}(Z) = \prod_{j=1}^{K} \frac{\Pi(Z_j - 1)}{\Pi(Z_j - 1) + Z}. \tag{50}$$

Equating the denominator of (16) to zero for S = Π (Z - 1), as Π → ∞, we obtain:

$$e^{S}\left(1 - \frac{pS}{K}\right)^{K} - 1 = 0. \tag{51}$$

With account for the roots of (51), we have for the generating function P_{Π} (Z):

$$\lim_{\Pi \to \infty} P_{\Pi}(Z) = \prod_{j=1}^{K} \frac{S_j}{S_j + Z}, \tag{52}$$

where S_j are the roots of (51) for Re (S) > 0.

$$\lim_{\Pi \to \infty} \frac{P_n}{\Pi} \sum_{\nu=1}^{K} \frac{\beta_\nu}{\left(1 + \frac{S_\nu}{\Pi}\right)^{n+1}}, \tag{53}$$

/42

where:

$$\beta_\nu = \frac{-\prod_{j=1}^{K} \frac{S_j}{\Pi}}{\frac{d}{dZ} \prod_{j=2}^{K} \left[\left(1 + \frac{S_j}{\Pi}\right) - Z\right]\Big|_{Z=\left(1 + \frac{S_\nu}{\Pi}\right)}},$$

$$\lim_{\Pi \to \infty} \frac{M\{N\}}{\Pi} = \sum_{j=1}^{K} S_j^{-1}, \tag{54}$$

$$\lim_{\Pi \to \infty} \frac{D\{N\}}{\Pi^2} = \sum_{i=1}^{K} S_j^{-2}. \tag{55}$$

b) $\Pi = \infty, \quad K = 1.$

$$\lim_{\Pi \to \infty} P_{\Pi}(Z) = \frac{S_1}{S_1 + Z}, \tag{56}$$

$$\lim_{\Pi \to \infty} \frac{P_n}{\Pi} = \frac{S_1}{\left(1 + \frac{S_1}{\Pi}\right)^{n+1}}, \tag{57}$$

$$\lim_{\Pi \to \infty} \frac{M\{N\}}{\Pi} = S_1^{-1}, \tag{58}$$

$$\lim \frac{D\{N\}}{\Pi^2} = S_1^{-2}. \tag{59}$$

c) $H = \infty, \quad K = \infty.$

$$\lim_{H\to\infty} P_p(Z) = e^{-\rho Z}, \tag{60}$$

$$\lim_{H\to\infty} \frac{P_n}{H} = \frac{1}{n!}\rho^n H^{n-1} e^{-\rho H}, \tag{61}$$

$$\lim_{H\to\infty} \frac{M\{N\}}{H} = \rho, \tag{62}$$

$$\lim_{H\to\infty} \frac{D\{N\}}{H^2} = 0. \tag{63}$$

Calculations were made of the functions $M\{N\}/H$ and $D\{N\}/H^2$ as a function of H for two values of the loading $\rho = 0.8$ and 0.9 (Figures 1, 2).

REFERENCES

1. Kantor, A. V., V. G. Timonin and Yu. S. Azarova. In the collection: Apparatura dlya kosmicheskikh issledovaniy (Equipment for Space Studies). Nauka Press, Moscow, 1972.

2. Krayzmer, A. P., D. A. Boradayev, A. I. Gutenmakher, B. P. Kuz'min and I. Ya. Smolyanskiy. Assotsiativnye zapominayushchiye ustroystva (Associative Memories). Energiya Press, Moscow, 1964.

3. Barucha-Rid, A. T. Elementy teorii markovskikh protsessov i ikh prilozheniya (Elements of Markov Chain Theory and Their Application). Nauka Press, Moscow, 1969.

4. Saaty, T. L. Elements of Queueing Theory with Applications. Sov. Radio Press, Moscow, 1966.

5. Bailey, N. T. J. Roy Statist. Soc. Ser. B, 1954.

DISPERSION SPACE RADIO LINKS

L. G. Sapogin and V. G. Sapogin

ABSTRACT. We examine the dispersion method of information reception and transmission with use of the dispersive medium in which the radio waves propagate as an optimal filter.

Highly directional ground-based antennas with large effective areas and cooled high-sensitivity receivers are used for communication with distant spacecraft. With the narrow antenna pattern and weak signal, spacecraft search becomes a complex problem because of the long time for establishing communication. The uplink communication problem is simpler than the downlink problem for the following reasons: /43

a) low onboard transmitter power; at the present time a transmitter power of 100 W can be considered the limit, since increase of the distance from the vehicle to the sun leads to marked reduction of the onboard energy supply;

b) limited directivity of the onboard antennas because of the severe requirements on spacecraft attitude stabilization in space;

c) limited ground antenna effective area;

d) the tremendous distances and inexorable law of squares lead to a situation in which the communication problem at the present time is resolved using all possible approaches (combined method). Further technology development requires strenous efforts and large expenditures in all these directions at the present time.

It is well known that the outer-space and near-Earth plasmas in which radio waves propagate have dispersion, which has always been considered a serious obstacle for transmission of wideband and FM signals, since they are distorted strongly because of dispersion [1]. Such dispersion-type signal distortions increase with distance, and impose limitations on the transmitted information spectrum width. For lunar distances, the dispersion limitations require signal spectrum width $\Delta f < 40$ MHz, and at distances of billions of kilometers the required bandwidth is on the order of tenths of a Hertz. The existing long-distance radio communication techniques have ignored the dispersion properties of the medium and have considered only its attenuation. The present study shows the possibility of using the dispersion property of the medium in which radio waves propagate as an optimal filter. For a given medium, the signal spectrum can be selected so that the signal harmonic components, traveling with different phase velocities because of dispersion, combine at the reception point in the required phase and create a local spatial region in which the peak signal power is very high. Here, the dispersion limitations will have a different nature, and the proposed dispersion method of information reception and transmission may lead to some progress in the questions of ultra-long-range space communication.

GENERAL QUESTIONS OF WAVE GROUP PROPAGATION IN DISPERSIVE MEDIA

First of all, we shall clarify how wave groups (signals) propagate in dispersive media. This question is not new, and has been examined in connection with study of electromagnetic wave propagation in feeders [2], plasma [3], amplification of light pulses

in radar [5, 6]. If we start from the problem of compressing a wave group into a narrow pulse [7] (in the limit a δ-pulse), we can formulate two problems:

1) finding the form of the matched wave group (spectrum, modulation law) which is compressed by a system with given dispersion characteristics into a narrow δ-pulse;

2) finding the required dispersion characteristics of the system from the given form of the signal which is compressed into a narrow δ-pulse. /44

The first problem is of primary interest for the creation of dispersion-type space radio links, and the solution of the second problem is analogous. We shall in the following assume that the dispersive system is linear, i.e., its characteristics are independent of wave group amplitude. In nonlinear systems, the question of wave group propagation becomes exceptionally complex, and has been examined in part in [8].

As is known [9], if a signal v (t) acts on the system input, the output signal will be

$$U(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} K(\omega, l) S_v(\omega) e^{i\omega t}\, d\omega, \tag{1}$$

where

$$S_v(\omega) = \int_{-\infty}^{\infty} v(t) e^{-i\omega t}\, dt$$

is the input signal spectrum; K (ω, l) is the transfer function of the linear system. A linear dispersive medium or system can be represented as a symmetric quadripole with two independent parameters:

$$K(\omega, l) = e^{-\alpha(\omega) l} e^{i\frac{\omega l}{v_f(\omega)}}, \tag{2}$$

where α (ω) is the coefficient of attenuation or amplification of the medium; v_f (ω) is the wave phase velocity in the medium (the

dispersion characteristic of the system); and l is the length of the considered segment of the medium.

We shall examine two different dispersive systems $K_c(\delta, l)$ and $K_p(\delta, l)$, and two wave groups: broad X (t) and narrow Y (t) with spectra $S_x(\delta)$ and $S_y(\delta)$, respectively. Assume that under the action of the broad group X (t) on the input of the system $K_c(\delta, l)$, there arises at the output the narrow group Y (t), and that under the action of the narrow group Y (t) on the input of the system $K_p(\delta, l)$, there arises at the output the broad group X (t) (reversibility). It is obvious that:

$$K_c(\omega, l) = \frac{S_y(\omega)}{S_x(\omega)}, \tag{3}$$

$$K_p(\omega, l) = \frac{S_x(\omega)}{S_y(\omega)}. \tag{4}$$

Hence, it follows:

$$K_c(\omega, l) = \frac{1}{K_p(\omega, l)} = \frac{K_p^*(\omega, l)}{|K_p(\omega, l)|^2}, \tag{5}$$

$$\alpha_c(\omega) = -\alpha_p(\omega), \tag{6}$$

$$v_{f_c}(\omega) = -v_{f_p}(\omega). \tag{7}$$

We see from (6) and (7) that the characteristics of the "narrowing" and "broadening" systems are mirror symmetric relative to the ω axis. The transformations performed on the wave group are not equivalent to reversibility of the propagation process (i.e., replacement of t by -t). While the system with normal dispersion and attenuation α broadens the narrow pulse, in the system with anomalous dispersion and amplification α, this broad pulse will be compressed into a narrow pulse. We note that all processes in systems with $\alpha \neq 0$ are irreversible. If the dispersion law is not altered, the sign of the time for the pulse changes. For example, if a linear FM pulse with time-increasing modulation frequency is subjected to compression, then, as a result of broadening of the narrow pulse, /45 we obtain a broad pulse with time-decreasing modulation frequency.

We note that (6) and (7) are valid for mutual transformation of a definite narrow pulse into a definite broad pulse and vice versa, but not for an arbitrary broad or narrow pulse. In this case, the same dispersive system may narrow one pulse and broaden the other.

FINDING THE SPECTRUM OF A WAVE GROUP COMPRESSED INTO A δ-PULSE IN DIFFERENT DISPERSIVE MEDIA

Assume that in one case a wave group δ (t) arrives at the input of the system K (ω, l) and is broadened into Y_1 (t), while in the other case the wave group X_1 (t) acts on the same system and is compressed by the system into the function δ (t - τ). Then we can write:

$$K(\omega, l) = \frac{\int_{-\infty}^{+\infty} Y_1(t) e^{-i\omega t}\, dt}{\int_{-\infty}^{+\infty} \delta(t) e^{-i\omega t}\, dt}, \tag{9}$$

$$K(\omega, l) = \frac{\int_{-\infty}^{+\infty} \delta(t-\tau) e^{-i\omega t}\, dt}{\int_{-\infty}^{+\infty} X_1(t) e^{-i\omega t}\, dt}. \tag{10}$$

After integrating and equating the right sides, we obtain:

$$\int_{-\infty}^{+\infty} Y_1(t) e^{-i\omega t}\, dt \int_{-\infty}^{+\infty} X_1(t) e^{-i\omega t}\, dt = e^{-i\omega\tau} \tag{11}$$

or the relation between the spectra:

$$S_{y_1}(\omega) = \frac{e^{-i\omega\tau}}{S_{x_1}(\omega)}. \tag{12}$$

Since the spectrum of the δ-function:

$$S_\delta(\omega) = \int_{-\infty}^{\infty} \delta(t) e^{-i\omega t}\, dt = 1. \tag{13}$$

by comparing (13) and (1), we obtain the condition under which a wave group in δ-function form is obtained at the output of the dispersive system:

$$K(\omega, l) S_y(\omega) = S_\delta(\omega) = 1, \quad K(\omega, l) = \frac{1}{S_y(\omega)}. \tag{14}$$

This formula makes it possible to find from the specific form of the dispersion characteristic the spectrum of the signal which is compressed into a δ-pulse and, conversely, to find for the case of a specific signal, which is compressed into a δ-function, the required dispersion characteristic of the medium. The exact solution of this problem for a specific dispersion law and wave group always involves considerable mathematical difficulty in calculating the Integral (1), which, as a rule, cannot be calculated exactly, and is evaluated approximately by the steepest descent method or on computers. Therefore, it is better to proceed in the opposite direction — make the mathematical study for approximate dispersion equations, when the Integral (1) can be calculated exactly.

For an approximate qualitative evaluation of the signals which can be compressed into a δ-pulse, we shall use the following simple analytic equations, obtained by graphical approximation of the actual dispersion characteristics in some frequency region. The linear normal dispersion: /46

$$v_f = c\left(1 - \frac{\omega}{\omega_1}\right), \quad 0 < \omega < \omega_1, \quad K(\omega, l) = \exp\left[-i\frac{\omega_1 l}{c}\left(\frac{1}{\frac{\omega_1}{\omega} - 1}\right)\right]. \tag{15}$$

The hyperbolic normal dispersion:

$$v_f = c\frac{\omega_1}{\omega}, \quad 0 < \omega < \infty, \quad K(\omega, l) = \exp\left(-\frac{\omega^2 l}{\omega_1}\right) \tag{16}$$

The parabolic normal dispersion:

$$v_f = c\frac{\omega_1^2}{\omega^2}, \quad 0 < \omega < \infty, \quad K(\omega, l) = \exp\left(-i\frac{\omega^3 l}{c\omega_1^2}\right) \tag{17}$$

Approximation of anomalous dispersions yields the same results, but with inverse law of frequency modulation within the wave group.

We find X (t) for the linear normal dispersion:

$$X(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left[i\frac{\omega_1 l}{c}\frac{1}{\left(1 - \frac{\omega_1}{\omega}\right)}\right] e^{i\omega t}\, d\omega. \tag{18}$$

The sought intergral can be tabulated. The result will be:

$$X(t) = \sqrt{\frac{\omega_1}{4\pi l}} \exp\left[-i\left(\frac{\pi}{4} + \frac{\omega_1 l}{2c}\right)\right] \exp\left[i\frac{\omega_1}{4}\left(\frac{l^2 - c^2t^2}{lt}\right)^2\right], \qquad (19)$$

i.e., if a wave group of very high amplitude, with carrier frequency varying in accordance with a definite law, arrives at the initial moment of time at the input of a system with linear normal dispersion, then a δ-pulse is obtained at the output. Let us examine the same case for hyperbolic normal dispersion:

$$X(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(i\frac{\omega^2 l}{c\omega_1}\right) e^{i\omega t}\, d\omega = \sqrt{\frac{c\omega_1}{4\pi l}} \exp\left(-i\frac{\pi}{4}\right) \exp\left(i\omega_1 \frac{ct^2}{l}\right). \qquad (20)$$

Any wave group of constant amplitude and linearly increasing modulation frequency is compressed into a δ-pulse. For parabolic normal dispersion the analysis is difficult, since the resulting integral cannot be evaluated. Evaluation by the steepest descent method yields the same result as for hyperbolic normal dispersion, but with nonlinear modulation law within the wave group. The estimates made above are very approximate because of the unrealizability of such dispersion equations for the condition $\alpha = 0$. Moreover, the δ-function yields an infinitely broad spectrum, while real wave groups cannot have such a spectrum. Therefore, it is necessary to solve the problem for an output signal in the form of a real Δ-pulse.

FINDING THE DISPERSION CHARACTERISTICS WHICH COMPRESS SIGNALS INTO A REAL Δ-PULSE

The correctly introduced real Δ-function for wave group representation must satisfy the Laplace equation and be sufficiently simple so that the resulting integrals can be evaluated. We take an auxiliary function $\gamma(t, p)$, for which $p > 0$ depends continuously on t, and in the limit takes the values: /47

$$\lim_{p\to 0}\gamma(t, p) = \begin{cases} \frac{1}{2} & \text{for } t > 0, \\ 0 & \text{for } t = 0, \\ -\frac{1}{2} & \text{for } t < 0. \end{cases} \qquad (21)$$

This requirement is satisfied by a function whose graph coincides with the function $y = 1/2 \operatorname{sgn} x$,

$$\gamma(t, p) = \frac{1}{\pi}\int_0^\infty e^{-p\omega}\frac{\sin \omega t}{\omega}\, d\omega = \frac{1}{\pi}\operatorname{arctg}\frac{t}{p}. \tag{22}$$

The derivative of γ (t, p) with respect to t, which we denote by Δ (t, p), will be:

$$\frac{\partial \gamma(t, p)}{\partial t} = \Delta(t, p) = \frac{1}{\pi}\int_0^\infty e^{-p\omega}\cos \omega t\, d\omega = \frac{1}{\pi}\frac{p}{p^2+\tau^2}. \tag{23}$$

It is obvious that

$$\lim_{p\to 0}\Delta(t, p) = \begin{cases} 0 & \text{for } t \neq 0, \\ \infty & \text{for } t = 0, \end{cases} \tag{24}$$

$$\lim_{p\to 0}\int_{-\infty}^{\infty}\Delta(t, p)\,dt = 1, \tag{25}$$

$$\delta(t) = \lim_{p\to 0}\Delta(t, p). \tag{26}$$

In the case of conventional integration of this function, the limit passage must be performed after calculating the integral. In other words, the small parameter must have lower order of smallness than the increment Δt. An integral with such a Δ-function denotes calculation of the limit of the sum as $\Delta t \to 0$, $p \to 0$ and $\Delta t/p \to 0$. Such integrals are improper integrals, and the function utilized is a generalized function.

Since the introduced function Δ (t, p) is the real part of the analytic function

$$\frac{1}{\pi z} = \frac{1}{\pi}\frac{1}{p - it}, \tag{27}$$

It is a solution of the Laplace equation and, consequently, can describe real wave groups. We find the spectrum of the function Δ (t, p):

$$S_{\Delta(t, p)}(\omega) = \int_{-\infty}^{\infty}\frac{1}{\pi}\frac{p}{p^2+\tau^2}e^{-p\omega t}\,dt = e^{-|p\omega|}. \tag{28}$$

Now it is not difficult to find the connection between the wave group and the transfer function of the system which compresses this group into Δ (t, p):

$$K(\omega, t)S(\omega) = e^{-|p\omega|}, \quad K(\omega, t) = \frac{e^{-|p\omega|}}{S(\omega)}. \tag{29}$$

As p → 0, the Equation (29) becomes (14). We note that the form of (29) changes very little with change of the technique for specifying Δ (t, p).

COMPRESSION OF LINEARLY FREQUENCY-MODULATED WAVE PACKET WITH GAUSSIAN ENVELOPE

The complex espression for the linearly frequency-modulated wave group with Gaussian envelope has the form

$$S(t) = \exp(-t^2/T^2)\exp[i(\omega_0 t + \gamma t^2)], \tag{30}$$

where T means half the pulse duration at the 0.37 maximal amplitude level. For a linear ω variation law, the instantaneous signal frequency takes negative values in some region. In this connection, we assume that: /48

$$|2\gamma| T \ll \omega_0 \tag{31}$$

and the region of negative instantaneous frequency values corresponds to negligibly small wave group amplitudes. We find the spectrum of such a wave group:

$$S(\omega) = \int_{-\infty}^{\infty} S(t) e^{-i\omega t}\, dt = \int_{-\infty}^{\infty} \exp -\left[\left(\frac{1}{T^2} - i\gamma\right) t^2 + i(\omega - \omega_0) t\right] dt. \tag{32}$$

We denote 1 - iγT = μ, and obtain:

$$S(\omega) = \frac{T}{\mu}\exp\left[-\frac{(\omega-\omega_0)^2 T^2}{4\mu}\right] \int_{-\infty}^{\infty} e^{-z^2}\, dz = \frac{T\sqrt{\pi}}{\mu}\exp\left[-\frac{(\omega-\omega_0)^2 T^2}{4\mu}\right]. \tag{33}$$

We find the modulus of the complex number μ:

$$m = |\mu^2| = \sqrt{1 + \gamma^2 T^4}.$$

Separating the real and imaginary parts in (33), we have:

$$S(\omega) = \frac{2\sqrt{\pi T^2}}{\Omega}\exp\left[-\frac{(\omega-\omega_0)^2}{\Omega^2}\right]\exp\left[i\left(\gamma T^2\frac{(\omega-\omega_0)^2}{\Omega^2} - \frac{1}{2}\operatorname{arctg}\gamma T^2\right)\right], \tag{34}$$

where $\Omega = \frac{2m}{T} = 2\frac{\sqrt{1+\gamma^2T^4}}{T}$ is the effective spectrum width. Consequently, the amplitude spectrum of such a pulse has Gaussian form and is concentrated near the center frequency ω_0. Using (29), we find the transfer function of the optimal dispersive medium:

$$K(\omega) = \exp\left[-\frac{(\omega-\omega_0)^2}{\Omega^2}\right]\exp\left\{-i\left[\gamma T^2\frac{(\omega-\omega_0)^2}{\Omega^2}-\frac{1}{2}\operatorname{arctg}\gamma T^2\right]\right\}, \qquad (35)$$

where $|K(\omega)| = \exp\left[-\frac{(\omega-\omega_0)^2}{\Omega^2}\right]$ is the amplitude-frequency characteristic of the medium and $B(\omega) = \exp\left\{-i\left[\gamma T^2\frac{(\omega-\omega_0)^2}{\Omega^2}-\frac{1}{2}\operatorname{arctg}\gamma T^2\right]\right\}$ is the dispersion characteristic of the medium.

If a linearly frequency-modulated wave group is passed through a medium with these parameters, at the exit from the medium we obtain an amplitude-modulated wave group of bell-shaped form, for which carrier frequency modulation is absent:

$$U_{ex}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)B(\omega)e^{i\omega t}d\omega = \sqrt{m}\,e^{-\frac{t^2}{(T/m)^2}}e^{i\omega_0 t},$$

Comparing (36) and (30), we can see that the output signal duration is reduced by m times and the amplitude is increased by $\sqrt{m}$ times in comparison with the input group, i.e., the pulse has been compressed by m times. This also follows from the energy conservation law.

The signal energy at the entrance to and exit from the dispersive medium is defined by the expressions:

$$E_{en} = \frac{1}{2\pi}\int_{-\infty}^{\infty}|S(\omega)|^2d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty}\sigma^2(\omega)d\omega,$$

$$E_{ex} = \frac{1}{2\pi}\int_{-\infty}^{\infty}|S(\omega)|^4d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty}\sigma^4(\omega)d\omega.$$

The signal compression factor is equal to the ratio of the entering wave group duration to the exiting wave group duration: /49

$$m = \frac{T_{en}}{T_{ex}} = \frac{T_{en}E_{en}^2}{E_{ex}} = \eta\, T_{en}\, F_{en},$$

where

$$\eta = \frac{\int_{-\infty}^{\infty} \sigma_0^2(\omega)\,d\omega}{\int_{-\infty}^{\infty} \gamma_0^4(\omega)\,d\omega}, \qquad \sigma_0(\omega) \quad \frac{\sigma(\omega)}{\sigma_m}$$

is the normed spectral density. Since $\sigma_0(\omega) << 1$, then $\eta \geq 1$. Thus, the compression factor is equal (in order of magnitude)to the product of the signal duration by its spectrum width:

$$m = \tau_{en} \cdot \Delta\omega.$$

For frequency modulated signals with large deviation frequency f_d, the occupied band width $\Delta\omega = 2\pi/f_d$, then:

$$m \simeq 2\pi f_d \tau_{en}, \tag{37}$$

but this is simply the modulation parameter of the frequency-modulated pulse. It is clear that for a given dispersive medium increase of the duration τ_{en} requires corresponding increase of the frequency deviation and this makes it possible, purely theoretically, to obtain any compression factors. In practice, the entire problem lies in radiating the frequency-modulated pulse with very large frequency deviation in accordance with a strictly defined law, since the stability of the medium at large distances is quite high.

ON DISPERSION OF MEDIA USED AS OPTIMAL FILTERS

According to the latest experimental data [3], interplanetary space is a plasma with particle concentration N = 100 el/cm^3. For such concentrations, we can neglect the magnetic field influence and consider the plasma nonmagnetic and collisionless. The dispersion equation of such a plasma has the form:

$$V_f = \frac{c}{\sqrt{1 - \frac{\omega_0^2}{\omega^2}}}; \qquad \omega_0 = \sqrt{\frac{4\pi e^2 N}{m}}. \tag{38}$$

For the interplanetary plasma, the frequency ω_0 is low, and for sufficiently high frequencies, we can expand (38) into a series

in powers of ω_0/ω, and consider only the first two terms of the expansion. Then,

$$V_\varphi = c\left(1 + \frac{1}{2}\frac{\omega_0^2}{\omega^2}\right), \tag{39}$$

which is greater than the speed of light. Formulas (38) and (39) lead to good agreement with the experimental data for the interplanetary and near-Earth plasma.

In the general case, the dispersion of outer space is determined by the magnitude of the integral electron concentration (IEC), which for near-Earth plasma varies continuously, depending on the time of day, season, and solar activity cycle in the limits $(1 - 8) \cdot 10^{13}$ el/cm^2. Change of the IEC may lead to change of the compression factor m (Figure 1). In addition, the magnitude of the IEC for near-Earth plasma depends on the magnitude of the zenith angle θ following the law (Figure 2): /50

$$N_1 = N_{1_0} \sec\theta,$$

where N_{1_0} is the IEC in the normal direction.

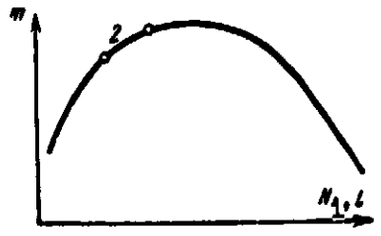


Figure 1. Compression factor as function of IEC and distance

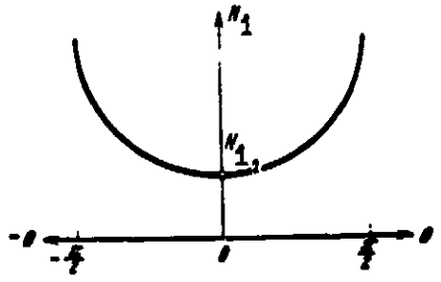


Figure 2. IEC as function of zenith angle

Studies made using artificial Earth satellites have shown conclusively that, along with the regular daily variation of N_1, there are observed irregular variations due to ionospheric nonhomogeneities. The ionospheric nonhomogeneity dimension spectrum occupies the

interval from several hundred meters to several hundred kilometers, and the magnitude of the electron concentration fluctuation increases with increase of nonhomogeneity size and lies in the range $\Delta N/N$ = 0.1 - 50%. Studies of the Mars and Venus ionospheres during flights of the Mariner 4 and Mariner 5 spacecraft make it possible to evaluate the magnitude of the dispersion of the ionospheres of these planets. Using the electron concentration distribution profiles in the Mars ionosphere [10] and Venus ionosphere [11], we can find the IEC along the radio wave propagation path and the rate of its variation as the vehicle approaches the planet. Calculations show that N_i along a ray in the Mars ionosphere may reach values of $N_i \sim 10^{13}$ el/cm^2, and in the Venus ionosphere — several orders greater, $N_i \sim 10^{15} - 10^{16}$ el/cm^2. According to the latest data, in the Sun's supercorona region (outer region of the corona at a distance greater than five sun radii) the electron concentration N decreases, following the law [13]:

$$N_i\ (R) = 1.1 \cdot 10^{23} R^{-2},$$

where R is measured in meters. For super-long distance spacecraft studying the depths of the Galaxy, the IEC may reach $5 \cdot 10^{21} - 5 \cdot 10^{23}$ el/cm^2 [14].

On the other hand, the question of electromagnetic wave dispersion existence in a vacuum has been posed many times, and is of definite theoretical and practical interest. Rozenberg's data on electromagnetic wave velocity measurement in a vacuum for a very wide frequency band after averaging yield the following table [15, 16]:

Wave frequency f, Hz	10^9	10^{14}	10^{19}	10^{22}
Propagation velocity c, km/sec	299,787.4	299,781.7	298,300	297,400

We see that the difference of the velocities at the edges of the band studied amounts to 1%, which corresponds to velocity change per unit frequency equal to: /51

$$T = \frac{\Delta c}{\Delta f} = 0{,}23874 \cdot 10^{-9}.$$

More precise measurements made in the RF band confirm this value of T. Rozenberg himself [16] questions the existence of dispersion in a vacuum, in contrast with other authors. Specifically, Teller [19] proposed exploding an atomic bomb in space and measuring the difference of the time of arrival at the Earth of the light and gamma rays.

It is not difficult to find the shift in seconds for frequencies of different electromagnetic waves emitted at the same time from some object:

$$\Delta t = t_1 - t_2 = \frac{L}{c_1} - \frac{L}{c_2} = \frac{L(c_2 - c_1)}{c_1 c_2} \simeq \frac{L\Delta c}{c^2} = L\frac{\Delta f}{c^2} T,$$

where c_1 and c_2 are the electromagnetic wave velocities at the frequencies f_1 and f_2. Kotelnikov and other authors [3, 12, 13, 17, 18] have drawn the conclusions on dispersion in interplanetary space.

DISPERSION LIMITATIONS IN LONG-RANGE SPACE RADIO LINKS

When organizing long-range space communications, it is necessary to consider the dispersion properties of the near-Earth plasma and interplanetary space, since signals are distorted markedly when transmitted over long distances. Therefore, it is important to clarify theoretically the limitations which arise. The group velocity of signal propagation in a dispersive medium is:

$$V_{gr} = cn,$$

where n is the refractivity of the medium. For a collisionless plasma,

$$n = \sqrt{1 - \frac{4\pi e^2 N}{m\omega^2}} = \sqrt{1 - \frac{f_0^2}{f^2}},$$

where $f_0 = 8 \cdot 10^7$; N is the number of electrons; e and m are the electron charge and mass. Then, the time of signal propagation

between the points X_1 and X_2 for arbitrary electron concentration variation law will be

$$t = \frac{1}{c}\int_{x_1}^{x_2} \frac{dx}{n} = \frac{1}{c}\int_{x_1}^{x_2} \left(1 - \frac{f_0^2}{f^2}\right)^{-\frac{1}{2}} dx,$$

where f is the signal center frequency. For $f^2 >> f_0^2$ (condition of radio transparency), this expression may be expanded into a power series, and we can write in the first approximation:

$$t = \frac{1}{c}\int_{x_1}^{x_2} dx + \frac{4 \cdot 10^7}{cf^2}\int_{x_1}^{x_2} N\,dx.$$

The first term is the signal propagation time in free space, and the second is the lag effect from dispersion, which depends on the frequency. Any transmitted signal occupies some frequency band Δf, which is connected with the signal duration τ by the known relation:

$$\Delta f = \frac{1}{\tau}. \tag{40}$$

Since the signal lag time is inversely proportional to frequency squared, we will observe a difference in its magnitude at the edges of the signal spectrum. Let the signal spectrum upper frequency be /52 $f_u = f + \frac{1}{2}\Delta f$, and the lower frequency $f_l = f - \frac{1}{2}\Delta f$. Then,

$$t_l - t_u = \Delta t = \frac{4 \cdot 10^7}{c}\left[\frac{1}{f_l^2} - \frac{1}{f_u^2}\right]\int_{x_1}^{x_2} N dx.$$

If $f \gg \frac{1}{2}\Delta f$, which is always the case, then

$$\Delta t = \frac{4 \cdot 10^7}{c}\left[\frac{(f_u - f_l)(f_u + f_l)}{f_l^2 f_u^2}\right]\int_{x_1}^{x_2} N dx = \frac{8 \cdot 10^7 \Delta f}{cf^3}\int_{x_1}^{x_2} N dx. \tag{41}$$

If the time Δt is commensurate with or greater than the transmitted signal duration, undetectable signal distortions and information loss occur. Expression (41) with account for (40) may be written in the form:

$$\Delta f < \sqrt{\frac{cf^3}{8 \cdot 10^7 \int_{x_1}^{x_2} N dx}}. \tag{42}$$

This inequality is the condition for undistorted signal transmission in long-range space radio links. For circumlunar distances [20], the quantity

$$\int_{x_1}^{x_2} N dx = N_1 = 5 \cdot 10^{18} \text{ el/cm}^2.$$

Distortions will not be observed for a signal frequency f = 1 GHz, if $\Delta f << 40$ MHz. For vehicles located beyond the Sun's supercorona or used for studying the depths of the Universe, the quantity Δf will be on the order of fractions of a Hz.

Changeover to the optical band (use of lasers) would make it possible to increase the signal spectrum width for long distances, but it is difficult to count on the appearance of such communications systems in the near future [21, 22].

INFLUENCE OF PLASMA NONHOMOGENEITIES ON WAVE GROUP PROPAGATION IN OUTER SPACE

Random nonhomogeneities influence wave propagation in a medium: they cause amplitude and phase fluctuations of the harmonic wave which depends on the frequency, i.e., there occurs a sort of dispersion which shows up in frequency dependence of the harmonic wave amplitude and phase fluctuation. Therefore, during propagation in a medium with random nonhomogeneities, the wave group will alter its form. Let us assume that weak large-scale nonhomogeneities are present only in the right-hand halfspace (x > 0). From the left-hand halfspace (x < 0), there is incident a wave group of arbitrary form:

$$U_0(x - ct) = \int_{-\infty}^{+\infty} C(\omega_1) \exp[i\omega_1(x - ct)]\, d\omega_1, \tag{43}$$

where the harmonic components of the wave group have the spectral amplitude

$$C(\omega_1) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} U_0(x - ct) \exp[-i\omega_1(x - ct)]\, d(x - ct). \tag{44}$$

In accordance with [23 - 25], the electromagnetic field intensity can be expressed in terms of fluctuations of the wave level B (k, x) and phase S (k, x): /53

$$U(\omega_1, x) = C(\omega_1)\exp[-\overline{B^2(\omega_1, x)} - iB(\omega_1, x)S(\omega_1, x) + B(\omega_1, x) + iS(\omega_1, x) + i\omega_1(x - ct)], \quad (45)$$

where C (ω_1) is defined by (44). After integrating (45) over all frequencies, we obtain the general expression for the field intensity in the wave group:

$$U = \int_{-\infty}^{\infty} C(\omega_1)\exp[-\overline{B^2(\omega_1)} - i\overline{B(\omega_1)S(\omega_1)} + B(\omega_1) + iS(\omega_1) + i\omega_1(x - ct)]\, d\omega_1, \quad (46)$$

hence,

$$UU^* = \iint_{-\infty}^{\infty} C(\omega_1)C(\omega_2)\exp\{-\overline{B^2(\omega_1)} - \overline{B^2(\omega_2)} - i[\overline{B(\omega_1)S(\omega_1)} - \overline{B(\omega_2)S(\omega_2)}] + B(\omega_1) + B(\omega_2) + i[S(\omega_1) - S(\omega_2)] + i(\omega_1 - \omega_2)(x - ct)\}\, d\omega_1 d\omega_2.$$

This quantity can be found as a result of statistical averaging of the factor

$$\exp\{B(\omega_1) + B(\omega_2) + i[S(\omega_1) - S(\omega_2)]\}$$

in the integrand. Considering a normal distribution law of the quantities $B(\omega_1)$, $B(\omega_2)$, $S(\omega_1)$, $S(\omega_2)$, it is not difficult to obtain the expression for the average value of this factor

$$\exp\left\{\frac{1}{2}\overline{B^2(\omega_1)} + \frac{1}{2}\overline{B^2(\omega_2)} + \overline{B(\omega_1)B(\omega_2)} - \frac{1}{2}\overline{S^2(\omega_1)} - \frac{1}{2}\overline{S^2(\omega_2)} + \overline{S(\omega_1)S(\omega_2)} + i[\overline{B(\omega_1)S(\omega_1)} + \overline{B(\omega_2)S(\omega_1)} - \overline{B(\omega_1)S(\omega_2)}]\right\},$$

then,

$$UU^* = \iint_{-\infty}^{\infty} C(\omega_1)C(\omega_2)\exp\left\{-\frac{1}{2}B^2(\omega_1) - \frac{1}{2}S^2(\omega_1) - \frac{1}{2}B^2(\omega_2) - \frac{1}{2}S^2(\omega_2) + B(\omega_1)B(\omega_2) + S(\omega_1)S(\omega_2) + i[B(\omega_2)S(\omega_1) + B(\omega_1)S(\omega_2) + (\omega_1 - \omega_2)(x - ct)]\right\} d\omega_1 d\omega_2. \quad (47)$$

We see that the mutual correlation and auto-correlation functions of the level B (ω) and phase S (ω) fluctuations at different frequencies appear in the integrand. If we assume that the correlation

coefficient for the refractivity fluctuation has Gaussian form:

$$n = \exp\left(-\frac{x^2 + y^2 + z^2}{L^2}\right),$$

with account for the method of [24, 25], we obtain

$$|\overline{U}^2| = \int_{-\infty}^{\infty}\int C(\omega_1)C(\omega_2)\exp\left\{-\frac{8\sqrt{\pi}x^2}{L^2}\left(\frac{\bar{\mu}_1^2}{D_1^2}+\frac{\bar{\mu}_2^2}{D_2^2}\right)+\right.$$
$$+\frac{32\sqrt{\pi}x^2\sqrt{\bar{\mu}_1^2\bar{\mu}_2^2}}{L^2D_1D_2}\cdot\frac{1}{D_1-D_2}\operatorname{arctg}\frac{D_1-D_2}{2}-$$
$$-i\frac{16\sqrt{\pi}x^2\sqrt{\bar{\mu}_1^2\bar{\mu}_2^2}}{L^2D_1D_2}\cdot\frac{1}{D_1-D_2}\ln\left[1+\frac{(D_1-D_2)^2}{4}\right]+$$
$$\left.+i\left[\psi(\omega_2)-\psi(\omega_1)+x(k_1-k_2)-(\omega_1-\omega_2)t\right]\right\}d\omega_1\,d\omega_2, \qquad (48)$$

where $\bar{\mu}_1^2$ is the mean square of the refractivity fluctuation $n_1 = n(\omega_1)$ at the frequency ω_1; $D_1 = 4x/k_1L^2$; the correlation function of the field fluctuation is assumed to be Gaussian, with nonhomogeneity dimensions L. We examine the propagation of an arbitrary pulse in a nonhomogeneous medium. We expand the exponent in (48) into a series in powers of $\nu_1 = \omega_1 - \omega_0$, and neglect terms of third order: /54

$$|\overline{U}^2| = \int_{-\infty}^{\infty}\int C(\omega_0+\nu_1)C(\omega_0+\nu_2)\exp\{-G(\nu_1-\nu_2)^2+$$
$$+i[Q(\nu_1^2-\nu_2^2)+P(\nu_1^2-\nu_2^2)]\}d\nu_1\,d\nu_2, \qquad (49)$$

where

$$G = \frac{\sqrt{\pi}xL\bar{\mu}_0^2}{8c^2}\left\{\frac{2}{3}D_0^2(n_0+\omega_0n_0')+\left[2(n_0+\omega_0n_0')+\omega_0n_0\frac{(\bar{\mu}_0^2)'}{\bar{\mu}_0^2}\right]^2\right\},$$
$$Q = \frac{1}{2}xk_0''+\frac{\sqrt{\pi}x^2\bar{\mu}_0^2}{Lc}\left[\frac{1}{2}(n_0+\omega_0n_0')\frac{(\bar{\mu}_0^2)'}{\bar{\mu}_0^2}+n_0'+\frac{1}{2}\omega_0n_0''\right]-\frac{1}{2}\psi_0'',$$
$$P = k_0'x-t+\frac{\sqrt{\pi}x^2\bar{\mu}_0^2}{Lc}(n_0+\omega_0n_0')-\psi_0'.$$

In these expressions, all the quantities with zero subscript relate to the frequency ω_0, and the differentiation is made with respect to ω. We shall examine an input signal with amplitude spectral density:

$$C(\omega) = C_0 e^{-\frac{(\omega-\omega_0)^2}{\Omega^2}} = C_0 e^{-\frac{v^2}{\Omega^2}}. \qquad (50)$$

This spectral density distribution selection makes it possible to examine all the physical phenomena of interest and, at the same time, simplifies the claculation of (49).

Substituting (50) into (49) and integrating, we obtain:

$$|\overline{U^2}| = \frac{\pi C_0^2}{\left[\left(\frac{1}{\Omega^2}+G\right)^2+Q^2-G^2\right]} \exp\left\{-\frac{P^2}{2\Omega^2\left[\left(\frac{1}{\Omega^2}+G\right)^2+Q^2-G^2\right]}\right\}. \qquad (51)$$

This expression describes the influence of plasma nonhomogeneities on wave group intensity. The nonhomogeneity factor is taken into account by the coefficient G. By definition, G depends linearly on distance. We see from analysis of the pre-exponential factor in (51) that the wave group intensity decreases inversely as distance because of the defocusing action of the nonhomogeneities. This conclusion is independent of wave group form and nonhomogeneity variation law. The only requirement is the normal nonhomogeneity distribution law, which, generally speaking, follows from the Lyapunov central limit theorem [26].

DISPERSION-TYPE SPACE RADIO LINK

When organizing optimal long-range space communications using the proposed method, it is necessary that the spacecraft remain constantly in the region of maximal signal power. To this end, the maximal power region is displaced to follow the spacecraft by selection of the radiated signal. Information transmission is accomplished using Pulse Code Modulation (PCM) methods. The dispersion-type space link consists of the transmitter with variable signal deviation frequency, the medium, and the receiver whose passband can be varied on command from the Earth and must increase with increase of the communications distance. The frequency-modulated transmitter pulse with decreasing modulation frequency has the duration τ, which increases with increase of the distance, while the signal center frequency f_0 does not change (Figure 3). The rate of change of the

/55

modulation frequency $\beta = 2\omega_d/\tau$ and the center frequency f_0 in this signal are constant at all times, while the spectrum width increases linearly with distance. The receiver in this radio link must be boradband, and special study should be devoted to its detailed analysis. The filter for the receiver is the space medium itself, while the receiving element may be a broadband maser with controllable band, followed by a detector and amplifier.

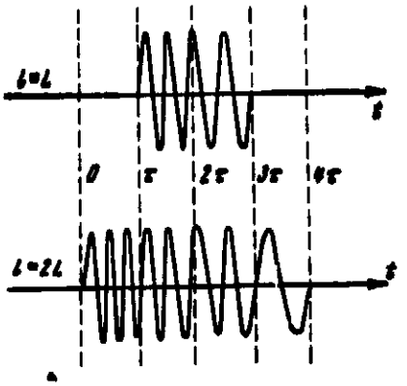


Figure 3. Types of radiated radio signals for different communication distances

After the video amplifier, the signal and noise enter a resolver with suitable observation criterion: this may be a Zigert-Kotel'nikov ideal observer, Neyman-Pearson minimal risk observer, or a minimax observer. For transmission of telemetry information from aboard the vehicle to the Earth, the average risk must be minimized by using group codes [27].

We shall analyze the variation of the signal compression factor with distance. With increase of the distance, the signal compression factor increases because of the dispersion of outer space, and decreases because of the defocusing influence of the plasma nonhomogeneities. If the compression factor decrease because of nonhomogeneities is greater than the factor increase, this will hinder organization of long-range space communications by the dispersion method. We have already seen that the signal compression factor is equal to:

$$m = 2\pi f_d \tau.$$

Substituting herein (41) and $f_d \simeq \Delta f$ (which is valid for large modulation factors), we obtain

$$m = \frac{16\pi \cdot 10^7 f_0^2 N_1 x}{c^3} - 1.$$

With increase of the distance, the required pulse duration will increase linearly. In order that the spacecraft be at all times in the maximal signal power region, it is necessary that the radiated signal duration and frequency deviation increase linearly with distance

$$F_d = AF_0x,$$

where F_0 is the initial frequency; A is a coefficient of proportionality; then,

$$m = \frac{16\pi A \cdot 10^{7} F_0 N_1 r^3}{cf^3}, \qquad (52)$$

i.e., the compression factor increases as the cube of the distance. With increase of the signal compression because of dispersion and defocusing by the nonhomogeneities, the signal power will increase as the square of the distance. Since the energy flux density at the reception point is inversely proportional to the square of the distance, we find that the peak signal power at the reception point will be independent of the distance. This conclusion is valid only in the case when, with increase of the distance to the spacecraft, the transmitted signal duration and frequency deviation increase and /56
the deviation magnitude is not too large.

In the case of compression of a rectangular frequency-modulated pulse, there arise additional side lobes on which signal energy is expended uselessly. Detailed examination of this question shows that the signal which is optimal from this viewpoint must have a Gaussian envelope. If we select a signal frequency deviation from 100 to 500 MHz, the maximal pulse duration for communications with Mars amounts to about 34 msec, and with Jupiter — about 0.5 sec. The compression factors obtained are $5 \cdot 10^7$ and 10^9, respectively. The dispersion-type space radio link can operate in two regimes: compensation for compression factor fluctuation by means of the plasma nonhomogeneities, and antenna "noise compensation" by varying the antenna zenith angles. The first regime is provided by selecting the operating point of the radio link at the maximum of the

compression factor versus IEC curve. As is known, antenna noise increases with increase of the zenith angle, which reduces the signal/noise ratio. On the other hand, the IEC also varies with zenith angle change. With selection of the radio link operating point in the region 2 (see Figure 1), we can obtain an approximately constant value of P_s/P_n for various zenith angles.

In order to establish contact with extraterrestrial civilizations and automatic interstellar stations of the future, it is necessary to resolve the problem of interstellar communications, which can be broken down into three basic aspects [14]:

1) is it possible to transmit a signal over very long distances and, if so, how is this best done;

2) how can we attract the attention of other civilizations, or vice versa;

3) how and in what "language" can meaningful information be interchanged with a completely foreign civilization.

The most important aspects are the first two, since the large distances make information transmission unusually difficult, and searching for signals at various frequencies and elevation angles is quite impractical. On the basis of the present level of the development of radiophysics, we can expect, in principle, the creation during the next 10 - 20 years of antennas with effective area 10^5 m^2, and receivers with noise temperature T_n = 1° K. Estimates for the isotropic case [28] show that such a receiver can record a signal from any point of the Universe from a transmitter with power $P \sim 10^{34}$ erg·sec^{-1}. At the present time, the total amount of energy radiated each day by mankind amounts to about $4 \cdot 10^{19}$ erg, while the Sun's energy output per second is $4 \cdot 10^{33}$ erg, which is an order of magnitude less than the required transmitter power.

The dispersion information transmission method makes possible a new approach to the problem of seeking extraterrestrial civilizations and providing interstellar communications. It is advisable to organize the search as follows. With the aid of phased antennas located at different points of the terrestrial sphere, we can create in the distant zone a spherical wave radiated by several frequency-modulated transmitters. By varying the transmitted signal duration and deviation magnitude, we can "probe" various spherical layers of the Galaxy, increasing the search radius in the course of time. In order that the signal be compressed into a δ-pulse in the region of UV-Cetus with frequency deviation from the optical to the RF band, a pulse duration of about 5 sec is required, and, in this case, the compression factor m $10^{12} - 10^{13}$. It is obvious that reception of extraterrestrial civilization signals must be accomplished with the aid of exceptionally wideband receivers of the Kotel'nikov receiver types. Only after establishing contact with the civilization is dispersion-type communications with directive antennas organized. The described technique of dispersion-type reception and transmission /57 is known in nature. Thus, certain forms of bats [29] apparently utilize the properties of acoustic dispersion of the surrounding air space and radiate ultrasonic frequency-modulated signals which are compressed into a δ-pulse as they propagate. This technique increases the range of action and resolution capacity of the ultrasonic radar.

In conclusion, we note that the described principles for the construction of dispersion-type space radio links provide potential noise immunity.

The authors wish to thank Prof. P. A. Agadzhanov, Dr. Tech. Sci., and Yu. K. Khodarev, Dr. Tech. Sci., for their valuable comments, discussions, and support in this study.

REFERENCES

1. Staras, H. Proc. IRE, Vol. 49, No. 7, 1961, p. 1211.

2. Collection: "Low-Loss Waveguide Transmission Lines:, ed. by V. B. Shteinshleiger. Foreign Literature Press, Moscow, 1960.

3. Ginzburg, V. L. Rasprostraneniye elektromagnitnykh voln v plazme (Electromagnetic Wave Propagation in Plasma). Nauka Press, Moscow, 1967.

4. Basov, N. G. and V. S. Letokhov. Dokl. AN SSSR, Vol. 1, 1967, p. 1966.

5. Shirman, Ya. D. Author's Certificate No. 146803, Bull. Izobr., No. 9, 1958.

6. Cook, C. E. Proc. IRE, Vol. 48, 1960, p. 310.

7. Sapogin, L. G., V. G. Sapogin and V. Ye. Lyamov. Trudy NIIP, Vol. 5 (117), 1969, p. 13.

8. Ostrovskiy, L. A. ZhTF, Vol. 33, 1963, p. 905.

9. Kharkevich, A. A. Bor'ba s pomekhami (Combatting Noise). Nauka Press Moscow, 1965.

10. Fjeldbo, G., W. C. Fjeldbo and V. R. Eshlemen. J. Geophys. Res., Vol. 71, 1966, p. 2307.

11. Mariner Stanford Group. Science, Vol. 158, No. 3809, 1967, p. 1678.

12. Kolosov, M. A., N. A. Armand and O. I. Yakovlev. Rasprostraneniye radiovoln pri kosmicheskoy svyazi (Propagation of Radiowaves in Space Communications). Svyaz' Press, Moscow, 1968.

13. Sobolev, V. V. Kurs teoreticheskoy astrofiziki (Theoretical Astrophysics). Nauka Press, Moscow, 1967.

14. Gindilis, L. M., S. A. Kaplan, et al. Vnezemnye tsivilizatsii (Problemy mezhzvezdnoy Svyazi) [Extraterrestrial Civilizations (Problem of Interstellar Communications)]. Nauka Press, Moscow, 1968.

15. Rozenberg, G. UFN, Vol. 48, No. 4, 1958, p. 599.

16. Rozenberg, G. UFN, No. 2, 1961, p. 349.

17. Kotel'nikov, V. A. Radiotekhnika i elektronika, Vol. 7, 1962, p. 1851.

18. Bunin, V. A. Astron. Zhurnal., Vol. 39, No. 4, 1972, p. 768.

19. Teller, J. Space Aeronautics, Vol. 6, 1959, p. 193.

20. Millman, G. H. J. Geophys. Res., Vol. 69, 1964, p. 429.

21. Ross, M. Laser Receivers. Mir Press, Moscow, 1969.

22. Chernyshev, V. N., A. G. Sheremet'yev and V. V. Kobzev. Lazery v systemakh svyazi (Lasers in Communications Systems). Svyaz' Press, Moscow, 1966.

23. Chernov, L. A. Rasprostraneniye voln v srede so sluchaynymi neodnorodnsotyami (Wave Propagation in Medium with Random Nonhomogeneities). Press of the Academy of Sciences of the USSR, 1958.

24. Bakhareva, M. F. Radiotekhnika i elektronika, Vol. 4, 1959, p. 88.

25. Shirokova, T. A. Akusticheskiy Zh., Vol. 5, No. 4, 1959, p. 485.

26. Hudson, D. Statistics for Physicists. Mir Press, Moscow, 1967.

27. Mitryayev, Ye. V. Radiotekhnika i elektronika, Vol. 8, No. 6, 1963, p. 923.

28. Kardashev, N. S. Astron. Zh., Vol. 41, No. 2, 1964.

29. McCue, J. J. G. International Convention Record, Pt. 6. Pergamon Press, New York, 1961, p. 310.

SPACECRAFT ANTENNA SYSTEM DESIGN

A. P. Alekseyev, B. A. Prigoda and
L. I. Skotnikov

ABSTRACT. We examine methods for ground development of unmanned spacecraft antennas. Considering the scope of this question, the accent is on the methods and devices for evaluating the characteristics of nondirectional antennas with ambient condition simulation which is as close as possible to the actual conditions. Special emphasis is given in the article to the questions of measuring directivity characteristics and studying antenna corona formation.

The problems of designing antenna systems having nearly optimal /58 characteristics are becoming increasingly important in connection with the rapid growth of space technology and, particularly, the development of unmanned space stations. In the present article, we examine some aspects of ground-based development of antenna systems associated with the methods and equipment for ground-based development with the closest possible simulation of the actual operating conditions of spacecraft of the subject types.

In contrast with the conventional antenna systems intended for stationary and mobile radio equipment designed to operate under ground-based conditions, the development of antenna and feeder systems (AFS) for spacecraft requires that several specific conditions

be met. Basically, these conditions reduce to the following two requirements: isolation from the influence of the Earth and surrounding objects; simulating the ambient medium conditions characteristic for the spacecraft during its operation. The validity of the results obtained during development of spacecraft AFS, and the quality and reliability of operation of the onboard radio equipment and of the spacecraft as a whole, will be determined by the degree to which these requirements are met.

Research and development associated with spacecraft antennas are usually carried out on special antenna test ranges, equipped with stands which permit measuring the spatial directivity characteristics, and chambers of various sorts in which the AFS are tested with partial or complete simulation of the actual ambient conditions. As a rule, either full-scale prototypes, the real spacecraft, or their models are tested. Antenna development is carried out on models constructed to some definite scale when full-scale operations are impossible because of the large size of the spacecraft, or because of the impossibility of avoiding the influence of the Earth and surrounding objects — for example, in the shortwave band and longwave part of the UHF band.

The capabilities of the modeling technique are limited to the cases when the structures of the antennas and of the spacecraft itself are relatively simple, and the external contours can be described by simple surfaces. In this case, modeling is accomplished comparatively easily and the electrodynamic properties of the model are very close to those provided by the surface of the actual spacecraft. However, in the case of complex spacecraft shapes, the creation of an electrodynamic model which simulates completely the surface impedance distribution of the actual vehicle is practically impossible. Therefore, the study of spacecraft AFS on a full-scale prototype is most acceptable. In this case, particular attention must be devoted to simulation of those components which are not structurally rigid, and may change their position or shape in the course of the flight. For example, this applies to evacuated

barrier thermal insulation (EBTI), which is a multilayer composition of fiberglass cloth and metallized film used to protect certain portions of the spacecraft surface, and also to the extendible booms, hatches, and so on. As a result of gaseous component release, the EBTI may expand in vacuum and its shape may change. Such changes can lead to redistribution of the surface currents excited by the antennas located on the spacecraft, which in turn causes distortion of the antenna pattern and disruption of communication between the spacecraft and the Earth. This effect is particularly undesirable if there are on board Doppler radio equipments, which are very sensi- /59
tive to parasitic fluctuation of the signal received by the receiver due to the presence of surface segments with varying impedance characteristics, which may lead to the appearance of a false signal and interruption in system operation.

Thus, the first and foremost requirement for AFS development is complete simulation of the external electrodynamic properties of the prototype, and assurance of the required stiffness of all the prototype components which participate in the formation of the basic parameters and characteristics of the spacecraft antenna system. In the following, we present some examples of how spacecraft AFS development is carried out. Because of the limited space available, the discussion is restricted to only certain AFS characteristics: spatial directivity pattern and electrical strength of the antennas under nearly realistic conditions. The study is made for broad-beam antennas, for which the spacecraft hull is an active element which influences these characteristics.

STUDY OF SPATIAL DIRECTIVITY CHARACTERISTICS

A large number of test facilities intended for the measurement of spacecraft antenna directivity characteristics have been described in the literature. In spite of their variety, they can be broken down into quite clear classifications. First, they can be subdivided on the basis of the degree of automation of the test object displacement process and recording and processing of the results into fully automated facilities, partially automated facilities,

and manually controlled facilities. Second, they can be classified on the basis of structural characteristics: with stationary measuring radiator, and with mobile radiator which displaces relative to the test vehicle center of mass. Other classification criteria can also be used, for example, based on the nature of the radiation source used: ground, airborne, or galactic sources.

In any case, regardless of the construction and nature of the test stand facilities used, they must be subject to the basic requirements on which the measurement accuracy and validity of the results obtained depend. These requirements amount to the fact that the distance between the test and auxiliary antennas must satisfy the distant-zone conditions: the signals reflected from the ground and foreign objects must not affect the measurement results, and the mutual coupling between the test and auxiliary antennas must be negligibly small. As is known from the literature, the first condition is satisfied when the distance R_{min} between the spacecraft being tested and the auxiliary antenna:

$$R_{min} = \frac{D^2 k}{2\lambda}, \quad (1)$$

where λ is the wavelength, D is the maximal antenna dimension; k is the coefficient defining the measurement error. If the tolerable error ΔE in determining the field intensity is no more than 1%, this expression takes the specific form:

$$R_{min} = \frac{2D^2}{\lambda}. \quad (2)$$

The measurement accuracy will increase with increase of the distance between the test and auxiliary antennas. This is shown in Figure 1.

In order to reduce the distortions caused by interference between the direct rays and those (secondary) rays reflected from the ground and local object, the area between the antennas must be open, the antennas should be located on towers, special reflecting panels or mats made from radiation absorbing material should be installed between the antennas. The antenna height above ground and the

/60

minimal acceptable distances from the nearest interfering objects which create secondary waves are determined from the formula:

$$H = \frac{\lambda}{2}\frac{R_{min}}{D_{rad}}, \qquad (3)$$

where D_{rad} is the linear dimension of the auxiliary antenna.

The antenna test stands are calibrated to exclude their influence on the test antenna characteristics.

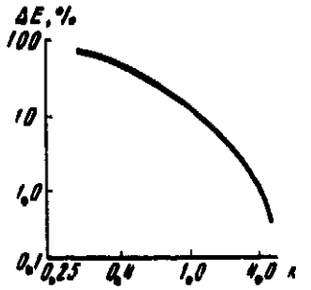


Figure 1. Measurement error versus distance between antennas

If the auxiliary antenna dimensions are less than those of the test antenna, i.e., the auxiliary antenna has a broad directivity pattern, then the Condition (3) becomes:

$$H \simeq \frac{D^2}{D_{rad}},$$

where D is the linear dimension of the test antenna. We see from what we have said that it is best to work with highly directional auxiliary antennas.

Figure 2 shows schematics of test stands used to study the directivity characteristics of spacecraft antennas. In the variant of Figure 2a, the test vehicle is mounted on a horizontal pivoting platform, which provides azimuthal rotation relative to the vehicle center of mass with any orientation of the vehicle axes in space. The variant of Figure 2b differs in that the vehicle is mounted on a system of flexible cables, rather than on a rigid base. By choice of the cable support system and using a drive located in the cable support system, it is possible to make measurements of the spatial directivity characteristics. This scheme is particularly convenient when conducting impedance measurements, and also in studies of the polarization characteristics, since it permits altering the vehicle height above ground in wide limits, and thereby provides optimal isolation from the ground. In order to achieve optimal isolation, it is best that the towers and area between them be covered by

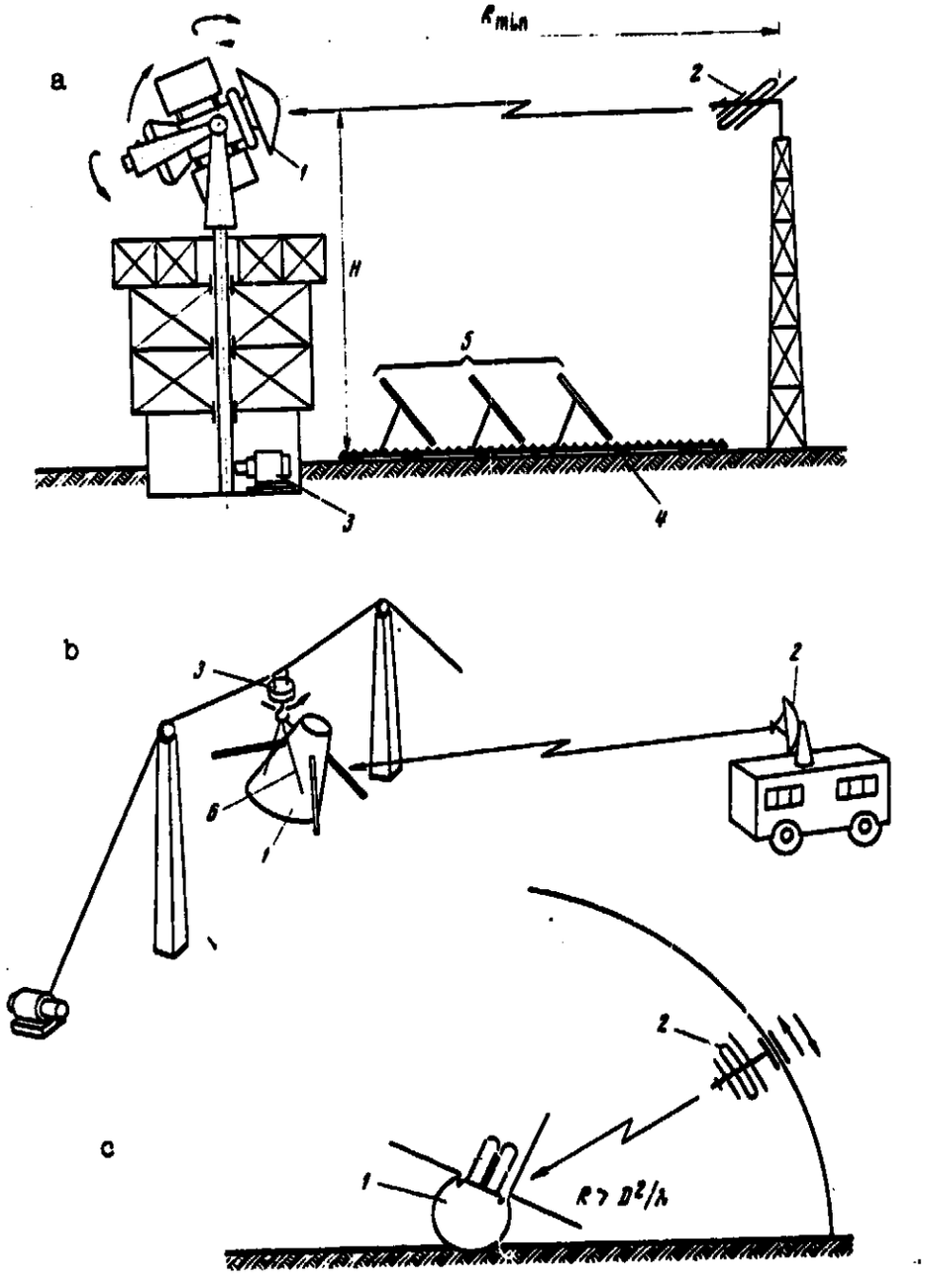


Figure 2. Antenna directivity characteristic measurement stands:

1 — test vehicle; 2 — auxiliary antenna; 3 — drive; 4 — radiation absorbing material; 5 — reflecting panels; 6 — cable support

panels and mats made from radiation absorbing material. The test stands in the Figure 2a and 2b variants are used in studying spacecraft antenna systems intended for operation under free-space conditions.

Special stands are used to conduct studies of the antenna systems of lander-type spacecraft which are intended for operation from the surface of planets. An example of such a stand is shown in Figure 2c. The landing vehicle is mounted on a surface simulating the assumed planetary soil model in the landing region. The auxiliary antenna displaces along a circular arc at a distance $R > D^2/\lambda$ from the lander's antennas. The signal level is recorded on the tape of a recorder which displaces synchronously with the measuring antenna movement. Studies of the directivity characteristics of the landing vehicles of the unmanned Luna, Mars, and other space stations were made in this way. /62

STUDY OF SPACECRAFT ANTENNA ELECTRICAL STRENGTH

One of the specific characteristics of spacecraft antenna operation is that of functioning under conditions of a highly rarefied gaseous medium. Observations made during flights at altitudes on the order of several tens of kilometers above the Earth have shown that antenna breakdown may occur even with low input power, measuring only a few Watts. Arcing occurs in those cases when the electrical field intensity reaches a definite level above the critical value. The use of dielectric coatings is advisable in order to reduce the field intensity level and reduce the possibility of arcing. The field intensity in the dielectric decreases in proportion to the magnitude of the relative dielectric permeability of the coating material.

If these measures are taken, breakdown of the gaseous medium surrounding the spacecraft will exert the primary influence on the electrical strength. Breakdown may show up either in the form of a spark discharge between individual parts of the structure, or as a corona discharge having the form of a plume which develops on some

part of the antenna. The nature and intensity of high-frequency breakdown are determined by the pressure and temperature of the gas, the presence of ionizing radiations, the frequency, average power, and duration of the high-frequency signal, the dimensions and shape of the antennas, and so on. As a rule, in studying the discharge mechanism and developing practical recommendations on compensating for the discharge, the primary attention is devoted to experimental rather than theoretical studies, since the latter constitute a quite formidable task.

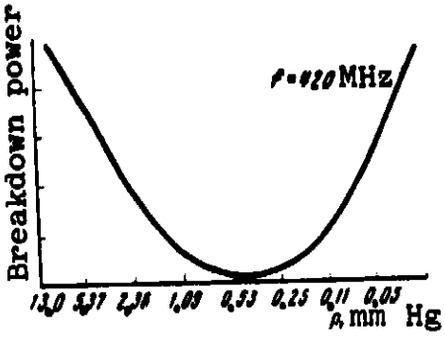


Figure 3. Breakdown power versus pressure

The minimal electrical strength of an air medium is obtained when the antenna electromagnetic field circular frequency ω is equal to the frequency ν of electron collision with neutral molecules, which is expressed in terms of the gas pressure p in mm Hg as follows:

$$\nu = 5{,}3 \cdot 10^9 p.$$

Consequently, we can write for $\omega = \nu$:

$$p\lambda = 36,$$

where λ is the wavelength in cm. Figure 3 shows the characteristic curve of breakdown power dependence on pressure for signal frequency 420 MHz. We see quite clearly the pressure region where breakdown shows up for minimal values of the power supplied to the antenna. With increase or decrease of the working frequency, this region will displace along the pressure scale to the right or left, respectively.

Special chambers are used to conduct such studies. Figure 4 shows a schematic of a chamber in which experimental studies were made of the whip antennas of the Mars 3 unmanned space station landing vehicle. The hemispherical fiberglass dome is equipped with viewing portholes. The vacuum pump system provides a vacuum of

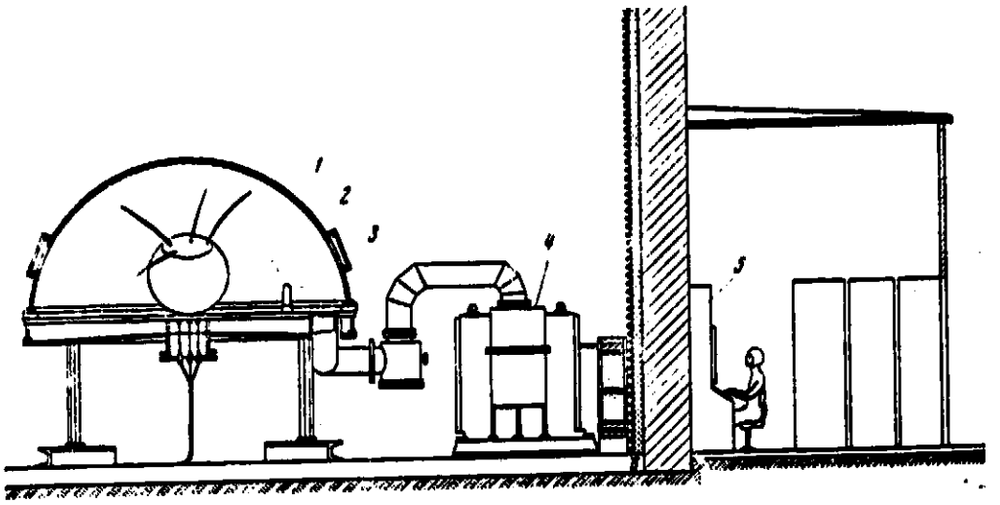


Figure 4. Radiotransparent vacuum chamber for developing antenna and feeder systems:

1 — vacuum chamber; 2 — test article; 3 — ionizer lamp; 4 — vacuum evacuation system; 5 — control console; 6 — radiation absorbing material

10^{-4} mm Hg, which is quite adequate for simulating the pressure both in the descent trajectory and after the landing vehicle touches down on the surface of the planet Mars. A special gas supply system makes it possible to create in the chamber cavity a gas composition simulating various Mars atmosphere models. Simulation of the solar radiation which creates the initial ionization was accomplished by tubes of the PRK (mobile X-ray unit) type located inside the dome. By altering the pressure, gas composition, and level of the power supplied to the antennas, we can evaluate quite completely the possibility of the occurrence of high-frequency breakdown on the antenna under foreign planetary atmosphere conditions. /63

Studies made in recent years have shown that gas composition has very little effect on breakdown level. The primary influence is that of pressure. The formation of ionization which transitions into corona is a very undesirable phenomenon which increases the high-frequency losses, changes the antenna input resistance, increases the standing wave ratio (SWR) in the channel, distorts the

directivity pattern, leads to partial or complete shielding of the antenna, failure of the onboard transmitter, and so on. The photographs in Figure 5 show how the corona formation distribution pattern on a whip antenna changes with change of the pressure of the medium in the chamber. Beginning at the end of the whip, at a pressure of 0.073 mm Hg, the corona formation extends over the entire length of the whip, reaching maximal intensity in the 1.0 mm Hg region for frequencies of $\sim$ 400 - 500 MHz.

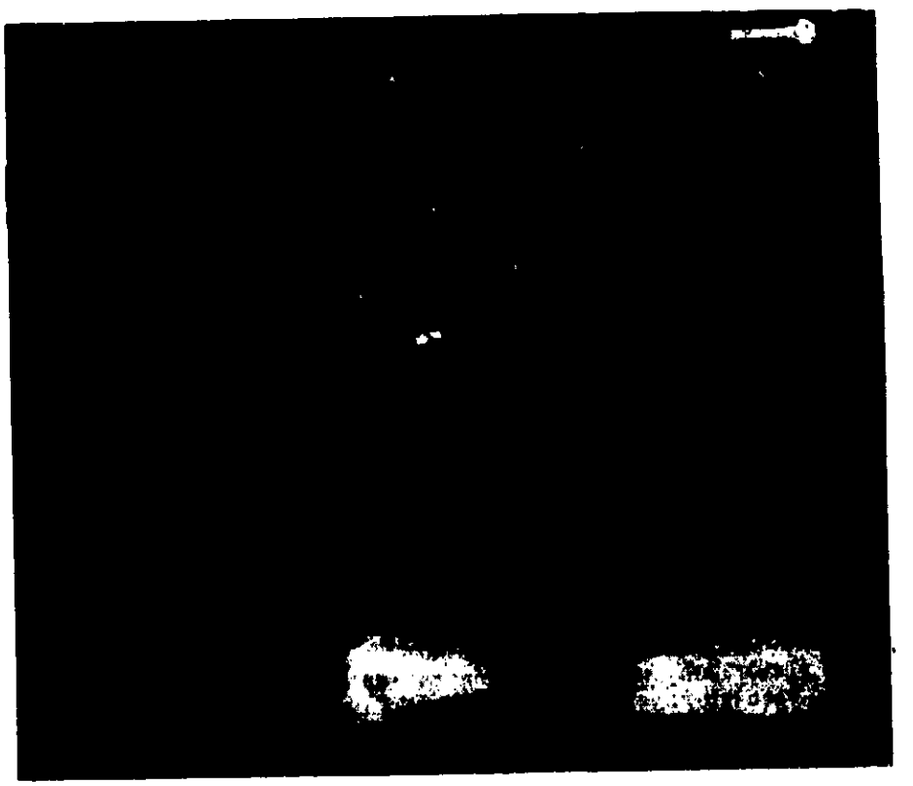


Figure 5. Corona formation on whip antenna at different pressures

It is obvious that the critical pressures in the atmospheres of the different planets do not show up at the same heights as for terrestrial conditions. Therefore, the problem of corona formation during entry into the atmospheres of these planets will arise in different parts of the entry trajectory. In this connection, it is important to select the transmitter frequencies and power of the capsules and landing vehicles intended for descent to the planet

surface, since the critical pressure at which breakdown begins is very sensitive to frequency. In the Earth's atmosphere, the point corresponding to the minimal breakdown voltage value appears in the region where the product of the pressure (in mm Hg) and wavelength (in cm) is between 10 and 40.

One of the problems in spacecraft antenna design is that of /64 gas release from the materials forming the spacecraft shell, antenna insulating washers, disks, and so on. An experiment has been described in the literature in which a slot antenna filled with polystyrene foam operated satisfactorily to pressures of 0.046 mm Hg with 200 W power. During repeat tests, after three-day exposure in a vacuum chamber, breakdown was observed at 80 W power. Gas release from the materials in high vacuum complicates this problem. The high temperature conditions on the surface of the Mars and Venus probes and landers make the problem associated with outgassing more critical.

INFLUENCE OF SURFACE HEATING AND PRESSURE

The influence of temperature and pressure at the spacecraft surface is important not only from the viewpoint of the possibility of distortion of the antenna characteristics as a result of corona formation. Heating of the spacecraft surface may lead to change of the properties of the dielectrics used to coat the antennas, change of the geometry of the antenna itself, and so on. In this case, there may be marked changes of characteristics, such as the antenna input impedance, which influences directly the channel SWR and radiated power level, and distribution of the high-frequency currents over the vehicle surface, which basically determines the spatial directivity pattern. It is obviously best to study the spacecraft or their individual components which are subject to heating or cooling, together with the antennas and high-frequency channel components which are located in these areas. Two measurement facility /65 schemes are used for this purpose. In the first case, the heat source and radio emission (radio reception) source are separated in space. Figure 6a shows a schematic of such a facility. The exhaust

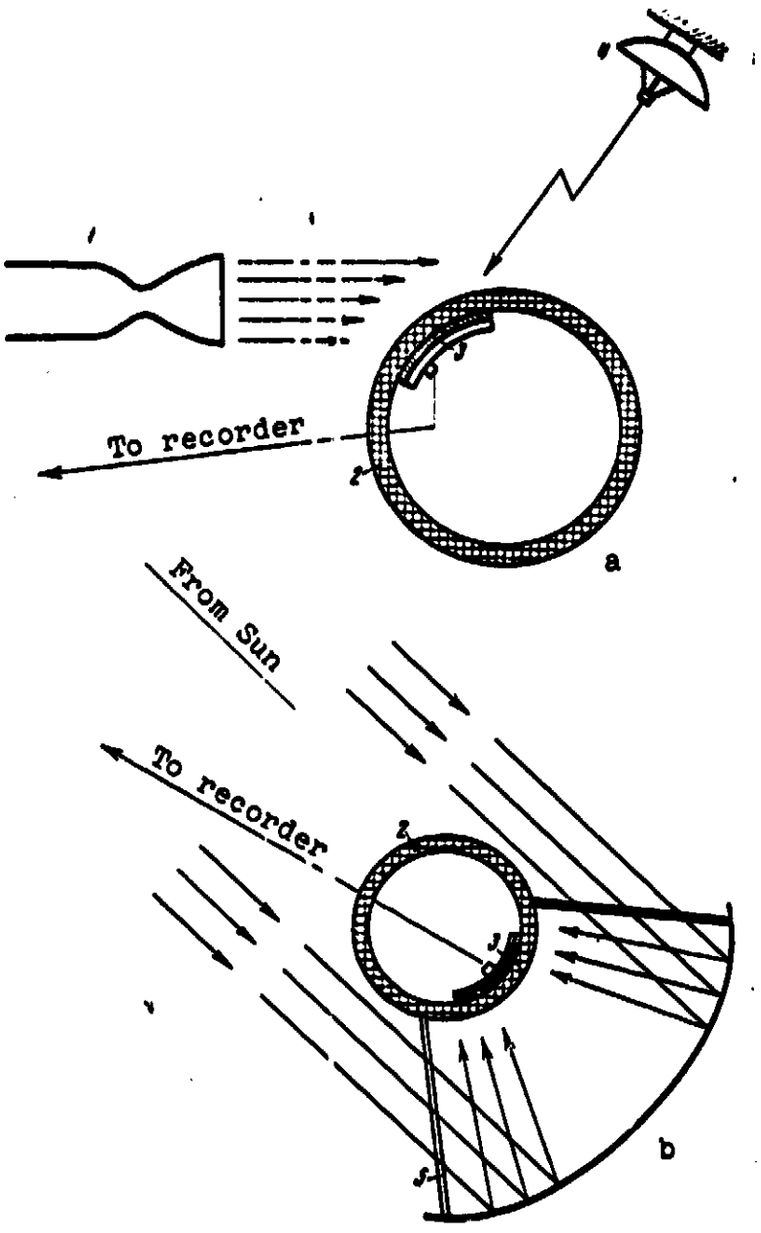


Figure 6. Schemes for heat testing lander antennas installed beneath thermal insulation layer:

1 — jet engine; 2 — lander thermal protection coating; 3 — test antenna; 4 — auxiliary antenna; 5 — mounting booms

of a jet engine is used as the heat source. The radio emission source is located at some distance from the test antenna. During the heating process, a record is made of the signal received by the test antenna, and the change of this signal, and also the SWR variation, is used to evaluate the degree of effect of heating on the /67

antenna. Another facility scheme (Figure 6b) is also possible, in which the Sun is used as the heat source and radio emission source. This is accomplished with the aid of a parabolic mirror concentrator at the focus of which the test vehicle with antenna is located. Heating is accomplished by the radiant energy of the solar spectrum and the Sun's electromagnetic flux in the RF band is used as the source of information on the nature of the change of the radiotechnical properties of the antenna. The advantage of this facility configuration is that the Sun's RF emission spectrum is practically continuous, and check of the antenna parameters can be made simultaneously over a very wide frequency range.

In conclusion, we shall discuss briefly the facility used to evaluate the operation of certain antenna units under high pressure conditions which occur, specifically, on the surface of the planet Venus. Such a facility is shown in Figure 7. It was used to test the Venus landers and, specifically, to develop the mechanism for firing the ejectable antenna of the landing module of the Venus 8 unmanned space station, which is shown in this same figure. The ejectable antenna was ejected at a gas pressure on the order of 100 atm, and heating of the ejection mechanism to 400 - 500° C. The effectiveness and reliability of the ejection mechanism was evaluated on the basis of the cage ejection height. The spacecraft antenna and feeder system can be equipped with various mechanisms, with the aid of which rotation, opening, and other displacements of the antenna feed system are accomplished. In these cases, AFS reliability must be verified on the Earth in order to increase the reliability of AFS mechanism operation under actual conditions. To this end, the spacecraft AFS mechanisms are tested in vacuum chambers and in thermostats, with heating or cooling of the mechanisms and the AFS.

Examination of the techniques for developing spacecraft AFS makes it possible to draw the following conclusions: test stand facilities intended for development of the spatial directivity

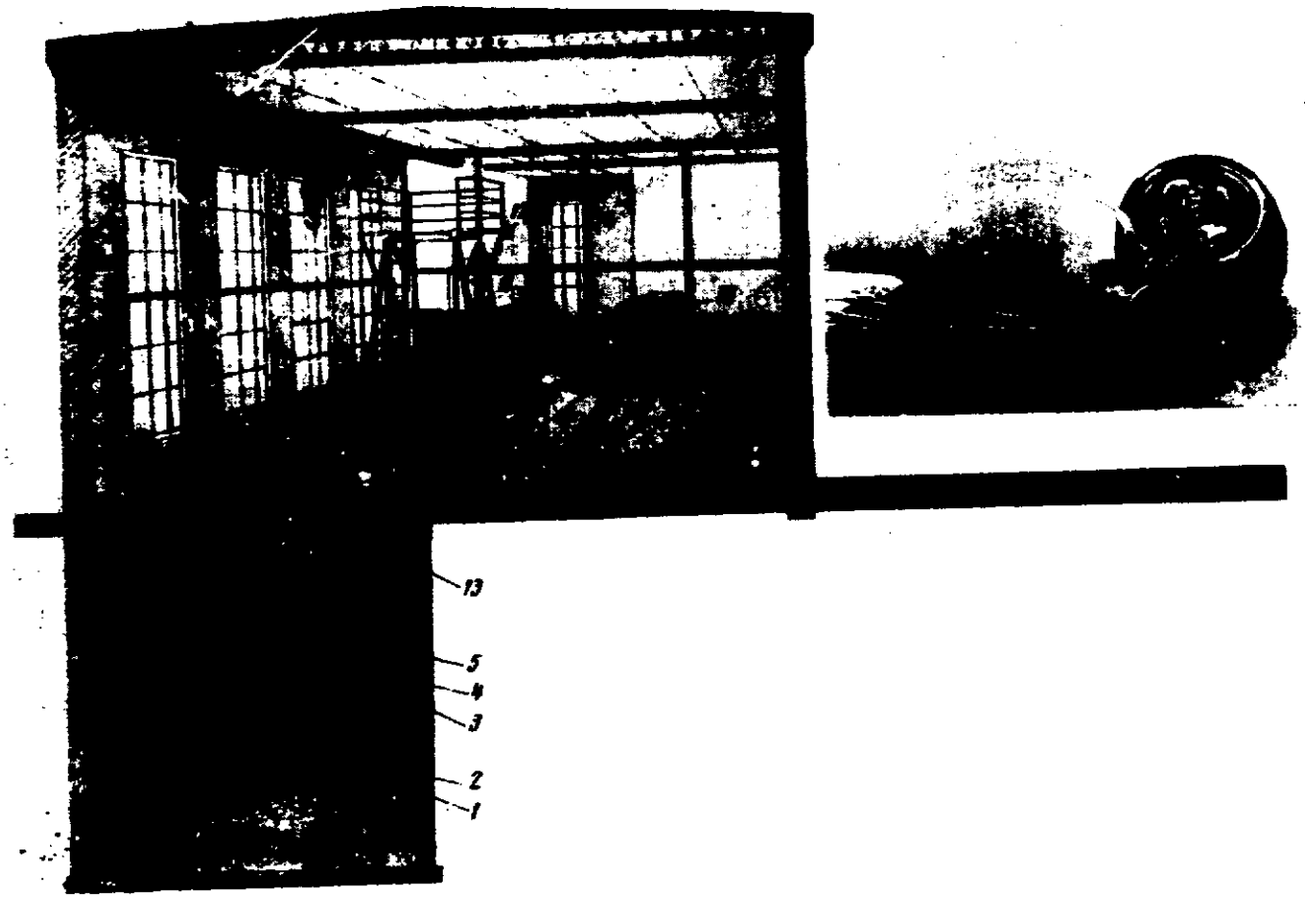


Figure 7. Stand for testing antenna ejection mechanism:

1 — antenna ejection mechanism; 2 — electric heater; 3 — movable calibrating weight with ratchet mechanism; 4 — guides; 5 — autoclave body; 6 — CO_2 gas generator; 7 — gas generator pneumatic console; 8 — autoclave pneumatic console; 9 — pneumatic system remote control pneumatic console; 10 — electric heater power supply panels; 11 — heater control panels and temperature measurement panels; 12 — autoclave cover handling stand; 13 — autoclave cover; 14 — electric power cables; 15 — electric instrumentation cables; 16 — CO_2 plumbing; 17 — electric hoist; 18 — instrumentation; I — Venus 8; II — ejectable antenna

characteristics of spacecraft antennas should provide maximal isolation from the Earth, and introduce minimal distortions into the measurements; the vacuum, thermal, and other special test stands must simulate, as far as possible, the actual conditions of spacecraft AFS operation: temperature, pressure, gas composition, Sun's radiation spectrum, and so on. The successful flights of the Luna, Mars, Venus spacecraft confirmed the effectiveness of the methods used for ground development of spacecraft AFS.

LOW-SILHOUETTE SPACECRAFT ANTENNA SYSTEMS

B. A. Prigoda

ABSTRACT. We examine ways to reduce the silhouette and overall dimensions of spacecraft antennas.

The modern spacecraft is connected with the external medium by /67
a system of sensors, optico-mechanical sensitive elements, antennas, and other special devices, each of which has a definite zone of action (spatial angle operating sector). In solving the general problem of studying outer space and the properties of the atmospheres and surfaces of the celestial bodies, we often must make compromises to provide the maximal capabilities for operation of a certain group of onboard devices and artificially narrow the zone of /68
action of the remaining instruments and devices, which are of less importance in the particular experiment. This approach is definitely undesirable, since it leads to a situation in which there is reduction of the potential capabilities of the scientific, telemetry, radio command and other systems, located both aboard the spacecraft and along the communication line between the spacecraft and the Earth. This frequently takes place because the zones of action of the various sensors, antennas, and other sensitive elements overlap spatially. When configuring these elements aboard the spacecraft within the limits of the specified overall weight and definite limitations on the dimensions, it is not always possible to separate

these instruments in space as they should be and mutual shadowing takes place, which leads to a situation in which — if there are aboard, for example, optical orientation sensors — certain parts of the active scan zone are cut off, and if the celestial body falls in these regions, interruptions will take place in astro-orientation system operation. This leads to lengthening of the orientation seances and, consequently, additional expenditure of the service life of the onboard equipment and power supplies, which in turn leads to increase of the overall weight, and so on.

In the present article, we examine some questions associated with reducing spacecraft antenna system dimensions, or, as we usually say in aviation, reduction of the antenna silhouette or design of low-silhouette antennas.

Up to ten or more antennas, each of which performs a definite task in regard to information reception or transmission, are installed aboard the modern spacecraft. The following approaches can be used in resolving the problem of reducing the antenna system overall silhouette and dimensions:

1) creation of frequency-independent antennas [1], i.e., combining the functions of several antennas operating in different bands and having analogous electrical characteristics into a single antenna unit;

2) creation of controllable antennas [2, 3], i.e., combining the function of several antennas with different electrical characteristics into a single antenna unit;

3) optimal configuration of the onboard antenna and feeder unit with scanning and connection of the antenna with optimal characteristics to a common feeder channel at each specific instant of time [4];

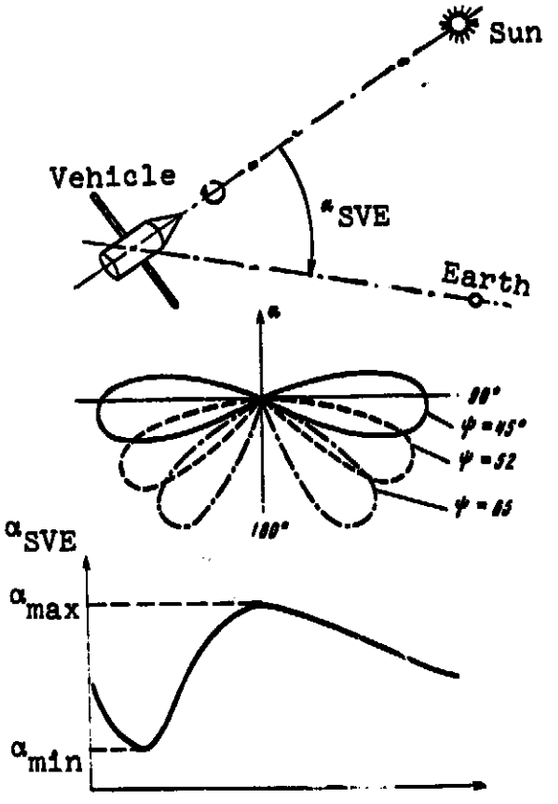
4) use of time diversity between the information reception and transmission seances in the case of overlap of the sensor spatial operating zones and antenna directivity patterns in the presence of their partial mutual shadowing;

5) use of magnetodielectric coatings which make it possible to obtain the required antenna characteristics with reduction of their overall dimensions [5].

The first technique has been quite thoroughly studied at the present time, and is widely used in practice. Thus, in spacecraft engineering various log-periodic structures (planar, cylindrical, hemispherical, and conical) are quite frequently and successfully used, which make it possible to obtain the same electrical characteristics in a wide frequency band with overlap on the order of 10 or more, i.e., $f_u/f_l \geq 10$, where f_l is the lower edge of the frequency band, f_u is the upper limit of the frequency band.

At the present time, the second technique is gradually beginning to be introduced into the engineering of spacecraft antenna units. This approach involves changing the configuration of the antenna or certain of its mechanical or electrical characteristics, so as to vary the required antenna parameters, depending on the specific task /69 imposed on the antenna at the given instant of time. The antenna system of the Venera unmanned space station can serve as an example. At certain times during the transfer trajectory, the station operates in the single-axis orientation regime, and rotates about the direction of the sun, close to or coinciding with the axis of principal moment of inertia. In order to provide communication with a ground station, it is obviously necessary, in this case, to have either an antenna unit which tracks the direction to the Earth or an antenna with radiation pattern of funnel-shaped form with maximum oriented toward the Earth. Since in the transfer trajectory, the Sun-vehicle-Earth (SVE) angle varies following a definite law, it is obvious that the Earth will gradually leave the zone of action of the antenna pattern maximum. In this case, in order to ensure

optimal communication along the entire transfer trajectory, it is necessary to have several antennas with funnel-shaped conical radiation pattern, but with spatially separated pattern maxima covering the zone of possible SVE angles for the given vehicle.



Curve of SVE angle variation and set of radiation patterns covering the SVE angle zone

The figure shows the curve of SVE angle variation and the set of radiation patterns covering the SVE angle variation zone ($\alpha_{SVE\ min} + \alpha_{SVE\ max}$). As the antennas which provide patterns of funnel-shaped form with elliptical polarization of the antenna radiation field, it is customary to use for spacecraft log-periodic two- and four-turn conical and hemispherical antennas for which the width and orientation of the radiation pattern maximum is determined basically by a single parameter — the spiral wrap angle. The radiations patterns shown in Figure 1 are provided by three four-turn conical log-spiral antennas with spiral wrap angles equal to 45°, 52°, and 65°, respectively. For the given antenna class, the problem of providing with a single antenna all three of the radiation patterns mentioned above reduces in practice to gradual variation of the wrap angle from 45° to 65°. A similar effect can be achieved by locating passive elements of the metallic disk or ring type at the apex of the spiral [3].

/70

The third and fourth techniques have much in common with one another, since they are based on alternate use of individual antenna elements with utilization of retractable booms, deploying mechanisms and other mechanisms which make it possible to reduce the dimensions of the system as a whole. In addition to the necessity for introducing certain mechanisms into the system, another drawback of these techniques is the presence of electronic analyzing devices, which are bound to have some influence on the reliability of the entire system. Studies are being carried out at the present time to optimize the onbaord system and maximize is reliability.

Reducing the size of the antenna radiators proper, even with account for the possibility of using the aforementioned techniques, is still advisable. Since simple reduction of the dimensions of any antenna leads directly to distortion of such parameters as the input impedance Z_{in}, directive gain (DG), radiation pattern shape, polarization properties of the antenna, and others, it is necessary to find techniques for reducing the geometric dimensions of the antenna such that its electrical characteristics will remain unchanged. One such technique is to locate the antenna in a magnetic-dielectric medium with parameters which differ from those of free space. For example, the above-mentioned multi-turn log-spiral concical antennas are remarkable in that their upper and lower working frequency band limits are determined by the truncated cone apex and base diameters. Since the wave propagation phase velocity, which defines its length in the given medium, depends on the properties ε and μ of the medium, we can assume that, other conditions being the same, the wavelength in the more dense medium ($\varepsilon > \varepsilon_0$, $\mu > \mu_0$) will differ from the wavelength in free space and, consequently, the effective dimensions of the radiator located in this medium will also differ. Studies which have been made show that the dimensions (a) of the active region of antennas coated with a layer of material with parameters differing from those of free space ($\varepsilon \equiv \varepsilon_0$, $\mu \neq \mu_0$), and of the uncoated antenna (located in free space) are connected by the expression:

$$n = \frac{a(\varepsilon, \mu)}{a(\varepsilon_0, \mu_0)} = \frac{\sqrt{\frac{1+\frac{1}{\mu}}{1+\varepsilon}}(1+\sin\psi\cos\theta_0)}{\left(1+\sqrt{\frac{1+\frac{1}{\mu}}{1+\varepsilon}}\right)\sin\psi\cos\theta_0}, \quad (1)$$

where ε is the dielectric permeability of the coating material; μ is the magnetic permeability of the coating material; ε_0, μ_0 are the corresponding parameters of free space; $2\theta_0$ is the spiral apex angle; ψ is the spiral wrap angle. Expression (1), obtained for conical spirals, can also be used successfully for bifilar spirals if we set $\theta = 0$. For example, for a bifilar spiral with wrap angle $\psi = 6.5°$, coated with a ferrite layer having $\varepsilon = 3.77$ and $\mu = 2.2$, the theoretical value of the size reduction coefficient is $n = 0.58$, in accordance with the formula shown above. This value of n is in good agreement with the values obtained in the laboratory in studying specimens of such antennas. Formula (1) shows that greater effectiveness can be achieved by using materials with large value of ε than materials with large μ for the coating. /71

The effective reduction of the antenna dimensions is determined not only by the properties of the coating material, but also by the coating thickness, its homogeneity, and certain properties of the specific antenna. The effectiveness of coating application will be different for different antenna types. The use of coatings with large value of ε or μ involves some undesirable effects, such as, for example, distortion of the input impedance, reduction of the channel traveling wave ratio (TWR), reduction of antenna efficiency, and so on, which must be considered during antenna development and compensated for in some definite fashion.

REFERENCES

1. IEEE Transactions on Antennas and Propagation, AP-13, N 3, 1965.

2. Prigoda, B. A.,et al. In the collection: Apparatura dlya kosmicheskikh issledovaniy (Equipment for Space Studies). Nauka Press, Moscow, 1972.

3. Mashkov, V. I. and B. A. Prigoda. In the collection: Apparatura dlya kosmicheskikh issledovaniy (Equipment for Space Studies). Nauka Press, Moscow, 1972.

4. Prigoda, A. B., M. B. Fainshteyn,et al. In the collection: Apparatura dlya kosmicheskikh issledovaniy (Equipment for Space Studies). Nauka Press, Moscow, 1972.

5. IEEE Transactions on Antennas and Propagation, AP-14, N 5, 1966.

HIGH-SENSITIVITY 3.5-cm MODULATION-TYPE RADIOMETER

A. Ye. Andriyevskiy, A. G. Gorshkov, V. V. Danilov, V. K. Konnikova, A. S. Lobarev, V. G. Mirovskiy, V. V. Nikitin, V. I. Portman, Ye. A. Spangenberg, I. A. Strukov, N. Z. Shvarts and V. S. Yetkin

ABSTRACT. A brief description of the circuit, characteristics of the functional elements, and parameters of a highly sensitive modulation-type direct-amplification radiometer operating in the 8000 - 9000 MHz band are presented. The radiometer was used together with the RT-22 antenna (Crimean Astrophysical Observatory of the Academy of Sciences of the USSR, located at Simenz). The sensitivity of the radiometric receiver was $\Delta T \leq 0.02^{\circ}$ K.

A high-sensitivity modulation-type radiometer, operating at the 3.5 cm (8.55 GHz) wavelength, was developed in 1967 - 1969 at the Institute of Space Studies of the Academy of Sciences of the USSR, the Moscow Pedagogical Institute im. V. I. Lenin, and the State Astronomical Institute im. G. K. Shternberg of Moscow State University, and has been used for radioastronomical observations on the antenna of the RT-22 Crimean Astrophysical Observatory (located at Simenz, Crimea). Extensive observational astronomical material has been accumulated, and the required information on the technical and operational parameters of the receiver has been obtained during the time of radiometer operation on the antenna. The results of the

/71

first stage of the observations were published in the Astronomical Circular of the Academy of Sciences of the USSR [1], are now being published in the Astronomical Journal of the Academy of Sciences of the USSR and in News of the CAO of the USSR, and will be published in the future as the data obtained are reduced. The objective of the present article is a brief technical survey of the components of the radiometric receiver.

The modulation radiometer is constructed using the direct amplification receiver scheme with $\sim$ 1000 MHz microwave circuit passband, and has sensitivity $\Delta T \sim 0.02^\circ$ K, with storage time constant $\tau = 1$ sec. Lobe modulation with lobe switching frequency F = 1000 Hz is used in the radiometer. A simplified block diagram of the radiometer's microwave circuit is shown in Figure 1. /73

Radiometer input circuits. The radiometer feeds provide operation in the lobe modulation regime and permit measuring the degree of circular polarization of the received radiation. Structurally, they are built in the form of two joined conical horns with circular waveguides (polarizers) and smooth transitions from the circular to rectangular section (Figure 2). The feed horns are located in the horizontal plane passing through the optical axis of the antenna system, and are displaced from the axis by $\pm$ 9'. The phase center of the feed horns is aligned with the focal plane of the antenna paraboloid. We note that for antennas with small aperture, the influence of attenuation fluctuations in the atmosphere on the radiation flux measurement accuracy decreases only slightly when using lobe modulation. For the RT-22 antenna, the sensitivity gain amounts to only a factor of four. But when using two horns, we obtain a characteristic output signal form which facilitates separating the source signal from noise when analyzing the output signal recordings.

The conical feed horns provide 15 dB irradiation intensity drop at the edge of the mirror. In this case, the antenna utilization factor is 45%; the first radiation pattern (RP) sidelobe has an intensity of less than 18 dB. The detailed nature of the side lobes was not investigated.

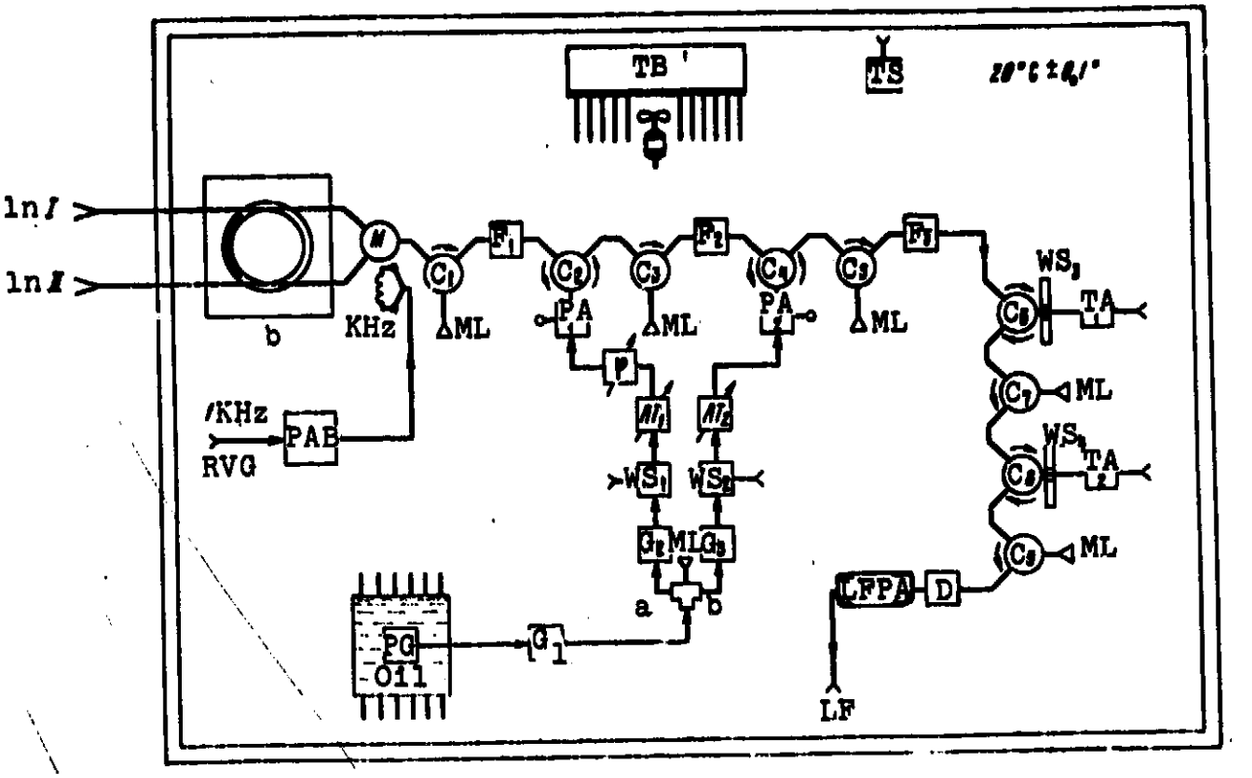


Figure 1. Block diagram of radiometer microwave circuit:

B — balancer; M — modulator; C — circulator; ML— matched load; F — filter; PA — parametric amplifier; TA — tunnel amplifier; WS — waveguide short; D — detector; PAB — power amplifier block; RFG — reference frequency generator; ϕ — phase shifter; AT — attenuator; G — gate; PG — pumping generator; TB — thermobattery; TS — temperature sensor

The device for measuring the degree of incoming radiation circular polarization consists of polystyrene quarter-wave dielectric plates ($\varepsilon = 2.5$), mounted in a circular waveguide, with one horn providing reception of radiation with right-hand circular polarization, and the other horn providing reception of

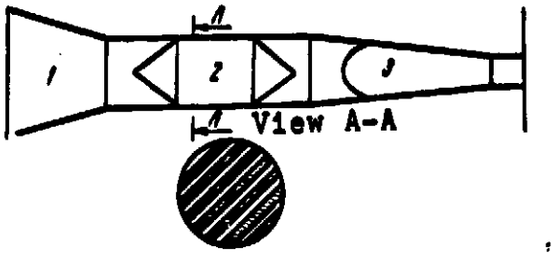


Figure 2. Schematic of antenna feed with polarizer:

1 — conical horn; 2 — polarizer; 3 — smooth transition

radiation with left-hand circular polarization. The difference in the record of the signal received by the first and second horns yields the degree of circular polarization and its sign. A drawback of the described polarizer is the difficulty in matching the polarizer with the receiver and feed circuit; as a result, it was not possible to obtain a feed SWR better than 1.7 in the band $\sim$ 1000 MHz. However, the quite high value of the feed SWR did not lead to marked deterioration of the fluctuation sensitivity of the radiometer, since the required isolation ($\sim$ 40 dB) was provided at the input of the first parametric amplifier. This feed system, together with the RT-22 antenna, made it possible to realize a radiation pattern of $(6.3' \pm 0.1) \times (6.6' \pm 0.1)$, with 18' angle between the main lobes.

Modulator. The modulator is a ferrite waveguide switch developed for this purpose in which change of the direction of circulation of the H_{10} linearly polarized electromagnetic wave is accomplished in a Y-circulator. A cylindrical insert made from type ZSCh-15 ferrite is used in the circulator. In order to reduce the eddy current losses in the circulator body and electromagnet magnetic structure (which is the basic problem in the design of a switch of this type if the switching frequency exceeds 150 - 200 Hz), the /74
waveguide tee was fabricated by the galvanoplastic method, and has waveguide wall thickness no more than 0.05 mm; the magnetic structure is fabricated from type 2000NM1 low-frequency ferrite, the solenoids are made in the form of two all-machined flat springs having 20 turns each. The entire modulator structure is pressed from plastic, and does not exceed in size the dimensions of a 3-cm band circulator. A /75
schematic of the switch construction is shown in Figure 3, and its electrical characteristics in the working frequency band are shown in Figure 4. A current amplifier is used to control the ferrite switch; the circuit provides relay commutation to block the switch in either of the extreme positions (control of microwave circuit amplitude-frequency characteristic, and so on).

Calibrator. The modulation radiometer with three-dimensional scanning of the antenna radiation pattern has a dual-channel microwave system ahead of the antenna switch (modulator). For

technological reasons, the components of the dual channel system cannot be made absolutely identical, and usually have attenuation scatter in the range of 0.1 - 0.2 dB. Due to this, the radiometer sensitivity decreases slightly (by 3 - 6%); however, an additional unbalance signal appears which is commensurate in magnitude with the antenna temperature increase during useful signal reception (up to 10 - 15°) from powerful cosmic RF sources. A special component is introduced into the microwave circuit to compensate for the loss unbalance in the two radiometer input channels. This component consists of two parallel rectangular waveguide segments into which an absorbing plate is inserted with the aid of an electric motor; this plate equalizes the losses in the microwave circuits. The maximum attenuation which

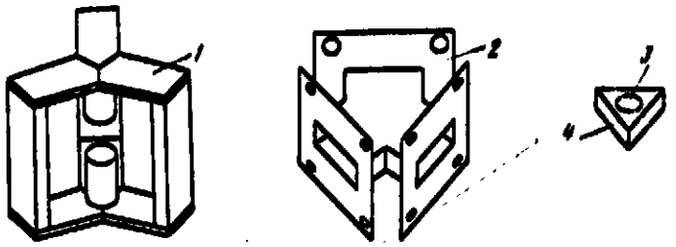


Figure 3. Modulator:
1 — magnetic structure; 2 — waveguide; tee; 3 — microwave ferrite; 4 — matcher

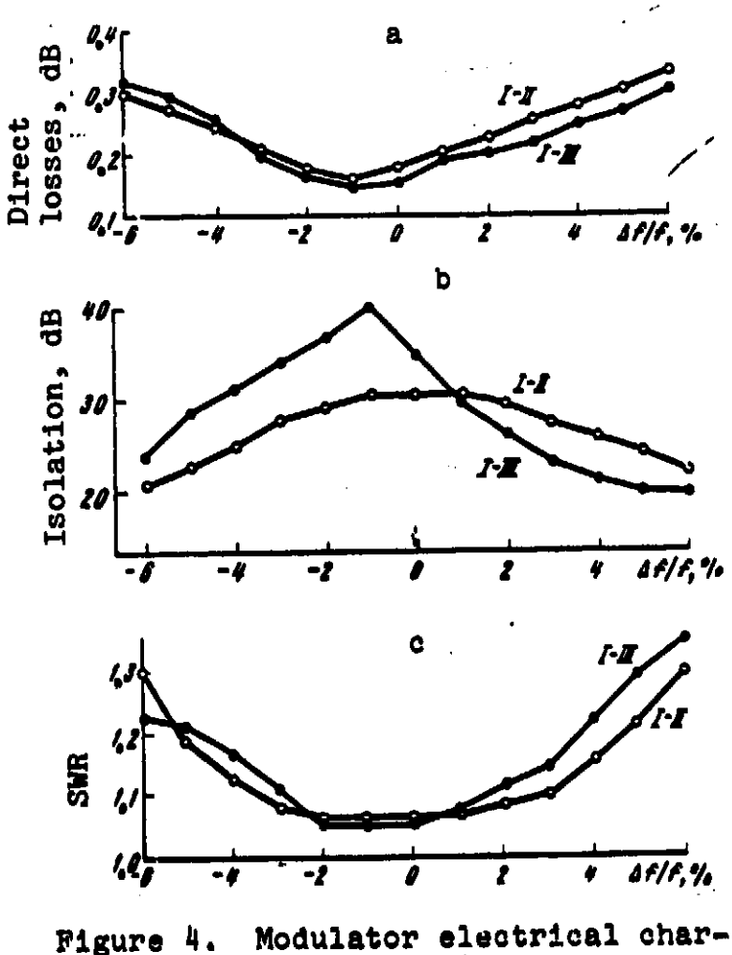


Figure 4. Modulator electrical characteristics:
1 — direct losses; 2 — isolation; 3 — SWR

can be introduced into either of the input circuit arms is ∿ 0.2 dB. On the common axis of rotation of the electric motor and the absorbing plate, there is mounted a type PTP-2 precision potentiometer from which an emf proportional to the attenuation introduced into one of the arms is taken. This emf is recorded by a pointer-type instrument on the radiometer control console, and makes it possible to calibrate the entire radiometer using an artificially introduced noise signal. An estimate of the accuracy of this calibration technique was made using the observational data, and the accuracy is about 3%.

The high-frequency amplifier block consists of two parametric and two tunnel amplifiers with wide passband 1000 MHz and overall gain G = 50 dB. An estimate of the required UHF gain was made following the calculation made by Korol'kov [2]. The sensitivity of the modulation type radiometer is defined by the expression:

$$\Delta T = \sqrt{2T_n^2 \frac{\Delta F}{\Delta f} + 4 \frac{T_0 \Delta F}{G^2 M^2 k \Delta f^2}}, \qquad (1)$$

where T_n is the receiver input equivalent noise temperature, Δf is the UHF circuit passband; ΔF is the LF circuit passband ($\Delta F = 1/4\ \tau$, where τ is the storage time constant); T_0 is the video detector temperature; G is the UHF circuit gain; k is the Boltzmann constant; M is the video detector quality factor. The first term of the radicand defines the fluctuations caused by the detected HF noise of all the receiver elements preceding the crystal detector; the second term defines the contribution to the fluctuations introduced by this detector and the subsequent LF stages. In a well constructed radiometric receiver, the condition must be met that the output noise introduced by the detector and LF section should be no more than 10% of the noise introduced from the microwave circuit:

$$2T_n^2 \frac{\Delta F}{\Delta f} = 10\left(4 \cdot \frac{T_0 \Delta F}{G^2 M^2 k \Delta f^2}\right), \qquad (2)$$

or

$$G = \frac{10}{T_n M} \sqrt{\frac{2T_0}{k \Delta f}}. \qquad (3)$$

In this case, the sensitivity of the compensation type radiometric receiver is determined basically by the first term of the radicand in (1): /76

$$\Delta T = \frac{1}{\sqrt{2}} \frac{T_n}{\sqrt{\Delta f \tau}}. \quad (4)$$

According to [2, 9, 10], the sensitivity of the modulation type radiometric receiver with degenerate parametric amplifiers at the input, rectangular modulation, and sinusoidal demodulation is

$$\Delta T \simeq \frac{\pi}{\sqrt{2}} \frac{T_n}{\sqrt{\Delta f \tau}}. \quad (5)$$

Repeated measurements of the sensitivity of the developed receiver under laboratory conditions and on the antenna (using calibration sources) showed good agreement of the experimental results with calculation using (5). The self-noise of the receiver without antenna did not exceed T_n = 150° K, and with individual parametric diodes decreased to T_n = 130° K. In connection with this low noise UHF level, we developed a special detector section with quality factor M = 150 - 200 $W^{-1/2}$. According to (3), with these basic radiometer element parameters, the optimal UHF gain should be G = 10^5 (50 dB). Further increase of the radiometer UHF gain is not advisable for the following reasons: 1) with increase of the UHF gain, the stability of regenerative UHF stage operation decreases; 2) the probability arises of saturation of the last tunnel amplifier stage; for $T_{n\,ant}$ = 30° K and gain G = 38 dB in the first three UHF stages, power P = $k (T_a + T_n) G\Delta f = 2 \cdot 10^{-8}$ W is applied to the input of the last tunnel amplifier.

The radiometer video detector maintains squareness of its characteristics reliably up to input signals not exceeding 10^{-6} W. High UHF gain will reduce the dynamic range of the radiometer. Therefore, when using the radiometer for observations of quite

sources (with antenna temperature $T_a \sim 500°$ K, provision must be made in such UHF schemes for disconnection of the last UHF stage (second tunnel amplifier).

Parametric amplifier block. Degenerate parametric amplifiers with gain G = 14 dB in the band Δf = 1 GHz are used as the first two UHF stages. The parametric amplifier passband lies in the range 8050 - 9050 MHz. The gain nonuniformity in the operating frequency range of each stage amounts to 0.2 - 0.3 dB. In the amplifiers, we used specially developed parametric diodes having short time constant, which makes it possible to realize quite low amplifier self-noise temperature. Thus, on some diode specimens, the parametric amplifier self-noise, measured at the input of the circulator and filter 1 (see Figure 1), amounts to $T_n \sim 90$ - 100° K. The circulators used had the required passband with isolation of 20 - 25 dB in each arm, and direct losses of $\sim$ 0.3 dB in the arm. We were able to select circulator specimens with low direct losses $\sim$ 0.18 - 0.2 dB in the arm for the radiometer input and for PA_1, which made it possible to reduce somewhat the receiver self-noise (by 7 - 15° K). The circulator C_1, providing the required isolation between the modulator and PA_1, was connected at the input (after the modulator M). This makes it possible to reduce the parasitic signal caused by interference of the amplifier self-noises with change of the modulator output impedance.

In order to prevent penetration of the pumping power from the PA to the modulator, which in turn may lead to the appearance of a parasitic signal, a LF filter (F_1) is provided at the input of PA_1 to reduce the pumping power by 40 - 50 dB. The losses at the signal frequency in the filter are quite small, and do not exceed 0.2 dB. /77
The filter is made with a two-sided "waffle" structure, and low losses are realized by maintaining high surface precision and surface finish. Prior to silvering, the filter surfaces were subjected to electrogalvanic polishing, which made it possible to improve

considerably the surface finish after machining. The filters F_2 and F_3 were fabricated similarly. The filter F_2 isolates the PA stages from the pumping power of the neighboring PA. The filter F_3 protects the tunnel amplifier block against the possibility of overloading by the pumping power which does leak through. The overall losses in the signal circuit from the input of circulator C_1 to the input of PA_1 (i.e., the arm C_1+ F_1 + the arm C_2) amount to about 0.45 dB.

Pumping circuit. The pumping generator is a klystron located in an oil bath to increase the heat rejection of the instrument. The pumping power from the klystron flows through the gate G_1 to the power divider, based on a dual T-bridge. The bridge was specially tuned to the pumping frequency with low active and reactive losses, and for optimal isolation of the arms in this case. As a result, VSWR parameters in the pumping generator arm equal to 1.05 were obtained; the isolation between the pumping arms of each PA is 14 dB; the pumping power flows along the two paths through the gate G_2, waveguide short WS_1, and attenuator AT_1 (similarly, through G_3, WS_2, AT_2) to the parametric amplifiers PA_1 and PA_2. The gates G_2 and G_3 isolate the PA stages at the pumping frequency, which simplifies tuning considerably and improves the operating stability of the stages. Gates with the following parameters were used in the pumping circuit: VSWR = 1.12, isolation 30 - 36 dB. The phase shifter ϕ is installed in the PA_1 circuit for selection of the optimal pumping phase. The VSWR of the parametric amplifiers in the pumping waveguide at the pumping frequency was 5 - 7. The electromechanical waveguide shorts WS_1 and WS_2 are used in the PA pumping circuits for convenience in stage-by-stage tuning of the PA, and for rapid check of operation of the radiometer stages.

Characteristics of degenerate PA cascading. In the described radiometer, both PA stages are supplied from a common pumping generator. As a result of this, the tuning and operation of the

parametric amplifier block have certain peculiarities: 1) the pumping phase of one of the cascades PA_1 must be optimized by means of the phase shifter ϕ in order that the overall gain of the two cascades be maximal throughout the entire amplification band for a given amplification level in each PA cascade; 2) the signal, after passing through the first PA_1 is modulated by the pumping frequency. The noise signal begins to carry "information" on the pumping phase and is processed by the second stage PA_2 as a synchronous signal. As a result, with optimal difference of the pumping phases on the two PA, the overall gain of the two stages is 3 dB higher than in the case when each of the amplifiers operates with its own pumping generator. This makes it possible to reduce the gain of the independently tuned PA to 11 dB and thereby improve somewhat the stability of operation of the radiometer as a whole. In this case, the overall gain of the two PA remains 28 dB.

Theoretical examination of the operation of degenerate parametric amplifiers in radiometers [2, 9, 10] shows that the interaction of the two bands, primary and unloaded, leads to reduction of the effective received frequency bandwidth by a factor of two. The expression (4) for the sensitivity of the compensation radiometer with degenerate input amplifier has the form

$$\Delta T = \sqrt{2}\frac{T_0}{\sqrt{\Delta f \tau}}$$

for sufficiently high regeneration. Here 2 Δf is the actual PA passband, observed with the aid of a sweep generator, i.e., in the expression for determining the radiometer sensitivity, there appears the factor 2, indicating deterioration of the sensitivity. This effect was confirmed experimentally by two methods. The first method involved connecting at the output of the parametric amplifier block a band filter with passband equal to one PA sideband — Δf (either primary or unloaded). Then the PA block passband was reduced by a factor of two in comparison with the sensitivity. The losses introduced by the filter were compensated in the signal circuit by a calibrating attenuator connected ahead of the detector. The

/78

measurements were made at a single detector working point (to avoid change of the transmission coefficient of the latter) and showed equality of the receiver sensitivity values with and without the filter.

The second method involved receiving the signal at the degenerate amplifier block output by a narrowband superheterodyne tunable receiver whose passband was on the order of 10 MHz, and was swept over the entire parametric amplifier band. One sideband of the superheterodyne receiver was suppressed by a special preselector. The measurement results were analogous to those of the first method. In both methods, the noise signal from the detector output was recorded on magnetic tape and then processed by an electronic computer. The expected qualitative difference was obtained in the probability density distribution functions of the noise signals processed by the degenerate and conventional parametric amplifiers [10, 12].

Tunnel amplifier block. The TA used in this radiometer has the following parameters: gain per cascade about 11 dB; passband $2\Delta f \sim 1$ GHz; noise temperature $T_n = 700 - 800°$ K. In the case of cascade connection of the two TA, the resulting $T_n \sim 800°$ K, which introduces at the radiometer input (with 28 dB gain of the two-cascade PA) less than 1° K excess noise. Circulators with waveguide cross section 5 × 23 mm were used as the TA block circulators. Electromechanical waveguide shorts (WS_3 and WS_4) are used for receiver tuning convenience, and when necessary for increasing the radiometer dynamic range, which is achieved by activation of the last stage TA_2.

Detector. We noted above that when designing highly sensitive radiometer receiver circuits, it is necessary to observe definite quantitative relationships between the detected microwave noise fluctuations and the low-frequency video detector noise fluctuations. It follows from (3) that for this purpose, a quite high microwave circuit gain (50 dB) is required in the direct-amplification radiometers. This leads to several undesirable effects.

Thus, because of the high gain in the UHF circuit in a single frequency band, additional screening and isolation are necessary in the microwave circuit, it becomes more difficult to broaden the UHF stage passband, the direct amplification receiver dynamic range decreases, and the stability of the radiometer as a whole decreases, which leads to deterioration of the receiver technical sensitivity. Increase of the video detector quality factor is a realistic way to decrease the gain in high-sensitivity radiometers. Therefore, in developing the radiometer primary attention was devoted to the problem of increasing the video detector sensitivity or, what is the same thing, increasing its quality factor. Among the diodes manufactured by Soviet industry, the highest quality factor value in the required band is provided by the D609 diode, which was used in constructing the radiometer detector section. The parameters of the detector section are shown in the table.

TABLE*

f, MHz	β, $\frac{\mu a}{\mu W}$	M, $W^{-1/2}$	P_{min}, W for Δf = 1 Hz	VSWR
7800	0,78	88	1,42	—
7,00	1,16	131	0,965	—
8000	1,62	183	0,695	2,85
8250	2,06	233	0,54	2,25
8500	1,98	224	0,61	2,04
8750	1,76	198,5	0,64	2,65
9000	1,52	176	0,72	4,3
9200	1,18	133,5	0,88	—
9400	0,98	110	1,15	—
9500	0,86	97,3	1,3	—

*Commas represent decimal points.

We find the quality factor and sensitivity from the formulas: /79

$$M = \frac{3R_v}{\sqrt{R_v + R_n}}, \quad P = \frac{\sqrt{4kT\Delta F}}{M}, \tag{6}$$

where R_v is the video resistance at the operating point; β is the anode current sensitivity; R_n is the video amplifier equivalent

noise resistance; P_{min} is the detector sensitivity; k is the Boltzmann constant; T_0 — 293° K; Δf is the video amplifier passband (equal to 200 Hz, when measuring the detector parameters).

Special selection of the D609 diodes is required in order to realize this high video detector sensitivity. Figure 5 shows typical curves of detector section quality factor versus frequency in the 8000 - 9000 MHz band for various D609 diode specimens. The high detector sensitivity required additional precautionary measures to reduce induction effects in the detector section and low frequency preamplifier circuits. Thus, in order to break the chassis currents, the detector section was galvanically isolated from the radiometer waveguide circuit and the coaxial cable between the detector and the LFPA was constructed with a high degree of shielding (double shield).

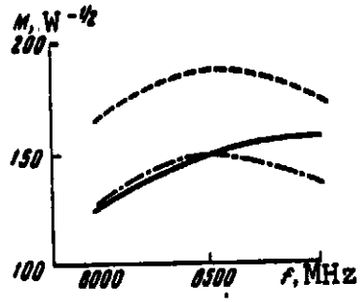


Figure 5. Detector section characteristics

Low frequency preamplifier (LFPA). The LFPA was located in the immediate vicinity of the detector section, and was constructed using two 6S51-N tube stages. The high input resistance and low equivalent noise resistance of the video amplifier provided maximal sensitivity of the detector-LFPA system. A filter which cuts off the power supply induction at 50 and 100 Hz was included in the amplifier circuit. The LFPA was placed in a double steel shield, and was damped mechanically with paralon (the latter made it possible to reduce by approximately an order of magnitude the microphonic effect of the mounting). The radiometer low-frequency block was the ShL-2 low frequency unit developed at the Special Design Bureau of the Institute of Radio Engineering and Electronics of the Academy of Sciences USSR with output to an EPP-09 recorder. One of the low frequency output channels was connected to an analog-digital converter, from the output of which information in binary code was fed to the magnetic recording system. /80

Radiometer microwave block. The entire microwave circuit of the radiometer was constructed in a single sealed thermostat, which was installed at the front focus of the RT-22 mirror. This same thermostat included the pumping generator (PG), power amplification block (PAB), power supply ferrite modulator, and the LFPA. Control of the microwave block components was accomplished by remote drives from the radiometer control console mounted in the RT-22 control room. The temperature inside the thermostat was maintained constant at +20° ± 0.2°, regardless of the ambient temperature, with the aid of a semiconductor thermobattery developed at the Semiconductor Institute of the Academy of Sciences USSR. A centrifugal fan was mounted in the thermostat to reduce the temperature gradient. Measurements showed that the overall gain drift of the entire microwave circuit of the radiometer together with the LFPA is $\Delta G = 0.5$ dB · deg^{-1}. The gain increases with increase of the temperature inside the thermostat.

In view of the fact that the microwave block was mounted on the moving portion of the RT-22 mirror, in constructing the radiometer it was necessary to increase the stiffness of the microwave circuit structure. Sealing of the microwave thermostat made it possible to eliminate the influence of marine climate on the radiometer components. The thermostat design made provision for the possibility of operation with a slight differential pressure of the nitrogen used to fill the chamber (0.01 - 0.05 at).

Operational characteristics of the radiometer. The equipment operated continuously without shutdown for several month-long operating cycles. During this time, the equipment operated very stably while retaining its basic technical parameters. The sensitivity of the radiometer on the antenna remained practically unchanged, $\Delta T \approx 0.02°$ K with $\tau = 1$ sec (confirmed by repeated recordings of calibrated sources). The instability of the gain of the entire radiometer channel was recorded in the laboratory, and did not exceed 1.3% during 30 minutes [a ("step") calibrated signal from a gas-discharge noise generator ($T_n = 25.5°$) was applied to the radiometer

input to check the gain stability]. The stability of the output indicator null decreased somewhat during operation of the equipment on the antenna because of the influence of atmosphere background fluctuations, instability of the power supply, and periodic radio noise at the RT-22 mirror location. On the average, the instability of the radiometer null caused by atmosphere background variations during good weather does not exceed 0.1° K (magnitude of the noise trace for the sensitivity of the given radiometer). In the case of moderate cumulus cloud cover, this instability lies in the range 0.1 - 0.25° K. During light rain, the instability reaches 1° K, which, for the RT-22 mirror corresponds to 1° $\approx$ 18 flux units, i.e., $18 \cdot 10^{-26}$ W/m^2 Hz. This large influence of the atmosphere

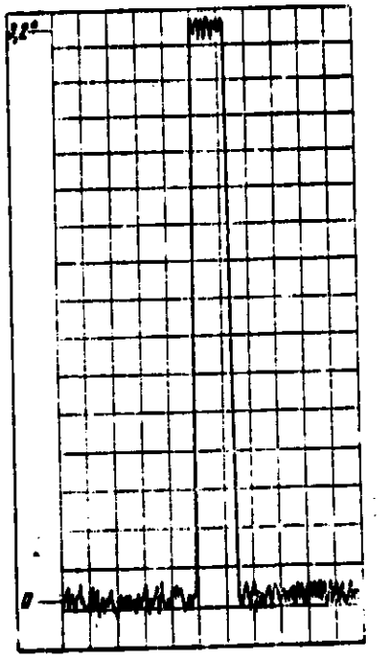


Figure 6. Calibration signal

/81

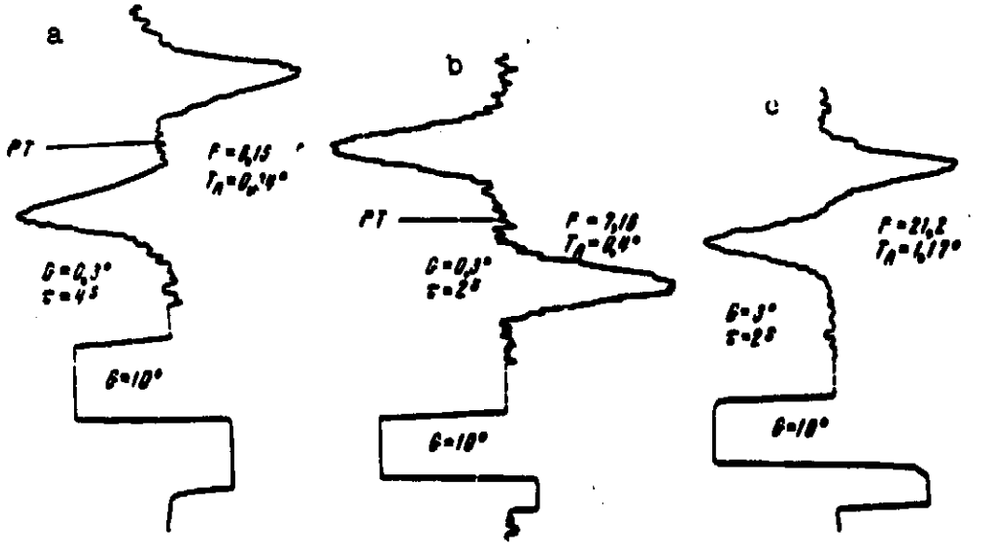


Figure 7. Source traces:
a — 30 - 218 (September 4, 1968); b — 4C - 3925 (September 4, 1968); c — 3C - 4543 (May 28, 1969)

on radiometer operation (even when using lobe modulation) is explained by the small magnitude of the RT-22 antenna aperture.

Figure 6 shows the laboratory recording of a 3.2° step with hot radiometer input 300° K, and for τ = 1 sec. The source shown in Figure 7c was recorded under unfavorable atmospheric conditions.

In conclusion, the authors wish to thank Academician A. B. Severnom and I. G. Moiseyev for the opportunity to use the RT-22 antenna for the observations, and also the personnel of the Crimean Astronomical Observatory of the Academy of Sciences USSR, the Institute of Space Studies of the Academy of Sciences USSR, the State Astronomical Institute im. Shternberg, and the Moscow State Pegagogical Institute im. V. I. Lenin for their active participation in the work using the telescope.

REFERENCES

1. Astronomical Circular of the Bureau of Astronomical Information of the Academy of Sciences USSR, No. 494, Feb. 21, 1969.

2. Yesepkina, N. A., D. V. Korol'kov and Yu. N. Pariiskiy. Radioteleskopy i radiometry SVCh (Microwave Radiotelescopes and Radiometers). Nauka Press, Moscow, 1973.

3. Korol'kov, D. V. and Yu. N. Pariiskiy. Izv. VUZov, seriya radiofizika, Vol. 11, 1968, p. 7.

4. Stankevich, K. S. Izv. VUZov, seriya radiofizika, Vol. 3, 1960, p. 765.

5. Ananov, N. I., A. Ye. Basharinov, K. P. Kirdyashev and B. G. Kutuza. Radiotekhnika i elektronika, Vol. 16, 1965, p. 1941.

6. Kidryashev, K. P. Radiotekhnika i elektronika, Vol. 12, 1967, p. 1487.

7. Tatarskiy, V. I. Rasprostraneniye voln v turbulentnoy atmosfere (Wave Propagation in a Turbulent Atmosphere). Nauka Press, Moscow, 1967.

8. Yetkin, V. S. and Ye. M. Gershenzon. Parametricheskiye sistemy SVCh na poluprovodnikovykh diodakh (Semiconductor Diode Parametric Microwave Systems). Soviet Radio Press, Moscow, 1960.

9. Markelov, V.A. Izv. VUZov, seriya radiofizika, Vol. 71, No. 3, 1964, p. 4?7.

10. Ivanenko, V. F. Radiotekhnika, Vol. 20, 1965, p. 14.

11. Kuril'chik, V. N., A. Ye. Andriyevskiy, V. N. Ivanov and Ye. Ye. Spangenberg. Astron. Zh., Vol. 46, 1969, p. 1124.

12. Spangenberg, Ye. Ye. and V. S. Yetkin. Izv. VUZov, seriya radiofizika, Vol. 10, 1967, p. 587.

IF AMPLIFIER LIMITING FREQUENCY SELECTION IN SUPERHETERODYNE MM- AND CM-BAND RADIOMETER

Yu. A. Nemlikher, I. A. Strukov and
L. H. Yudina

ABSTRACT. A technique is proposed for calculating the limiting frequencies of an intermediate-frequency amplifier from the experimentally obtained frequency dependences of the klystron generator noise radiation spectral density and IF amplifier temperature. An experimental relation and technique for measuring the amplitude-modulated noise spectral density of some klystron types are presented.

The thermal radar method using equipment installed aboard airplanes and artificial Earth satellites [1, 2] is being used more and more extensively for studying the natural resources of the Earth. Such equipment includes the radiometer with high fluctuation sensitivity ΔT, whose magnitude can be calculated using the known formula: /82

$$\Delta T = \frac{DT}{\sqrt{\Delta f \tau}}, \qquad (1)$$

where D is a coefficient which depends on the radiometer constructional characteristics; T is the receiver noise temperature, including the antenna noise; τ is the integrator time constant; Δf is the receiver passband up to the first detector. We find from (1 that to realize high radiometer sensitivity, it is necessary to reduce the receiving system input noise temperature T and broaden its passband up to the first detector.

In principle, two schemes for the construction of a radiometer for the millimeter and centimeter wavelength bands satisfying these requirements are possible: a) the direct amplification radiometer scheme; b) the superheterodyne radiometer scheme with shift of the signal spectrum into the region of lower (intermediate) frequencies. For all its apparent simplicity, the realization of the first radiometer scheme in the subject band encounters difficulties with operating frequency increase. With the appearance of boardband low-noise transistor amplifiers and resistive transformers using (GaAs) Schottky barrier diodes, it is becoming possible to obtain simple and highly sensitive input reception devices, which makes radiometers of the second type preferable [3].

The fluctuation sensitivity of the superheterodyne radiometer will be maximal only with proper choice of the IF amplifier working frequencies. The lower amplifier frequency boundary is determined by heterodyne noise and the upper limiting frequency is determined by IF amplifier (IFA) noise. In the general case, the noise temperature of the superheterodyne receiver, whose block diagram is shown in Figure 1, can be written in the form:

$$T = L_0 T_0 \left[\left(\frac{A}{f} \right)^m + \frac{t_a}{L_0} + t_0 + F_2 - 1 \right], \qquad (2)$$

where A is the coefficient characterizing the overall noise of the heterodyne klystron in the given regime, which is numerically equal to the intermediate frequency value at which the increase of the receiver optimal noise temperature, owing to heterodyne noise, is equal to unity; m is any positive number $m > 0$; f is the instantaneous frequency, lying in the low-noise IFA passband; t_a is the antenna relative noise temperature $t_a = T_a/T_0$, where $T_0 \approx 290°$ K; L_0 are the mixer power conversion losses; t_0 is the relative noise temperature of the mixer itself; F_2 is the IFA noise coefficient.

In order to find the average value of the receiver noise temperature in the IFA band Δf (Figure 2), it is necessary to know the form of the approximating curve for $F_2 - 1$, under the condition that /83

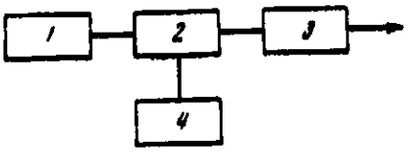


Figure 1. Block diagram of superheterodyne receiver input section: 1 — antenna; 2 — resistive frequency converter; 3 — wideband IF amplifier; 4 — klystron heterodyne

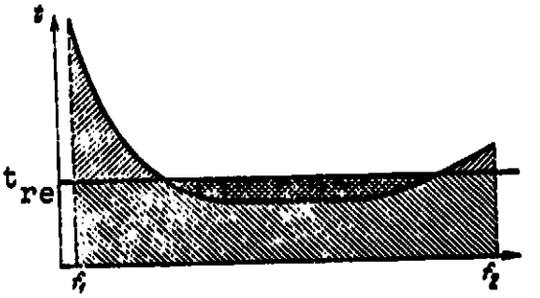


Figure 2. Illustration for finding average value of equivalent relative noise temperature

the mixer conversion losses L_0 are constant in the subject frequency band Δf. Experiment shows that the approximating function can be a power-law function:

$$F_2 - 1 = \left(\frac{f}{B}\right)^k + t_2, \qquad (3)$$

where the coefficient B and exponent k characterize the noise properties of the transistor IFA, beginning at some frequency. Provided that $f \ll B$, the amplifier noise coefficient is constant and equal to $F_2 = 1 + t_2$. This is correct, since the IFA lower limiting frequency is bounded by the heterodyne noise, and is selected to be no less than 30 - 50 MHz. The coefficient B and exponent k are determined not only by the transistors used in the IFA, but also by the number of amplifier stages, the amplifier circuit solutions, and so on. For example, for one of the amplifiers using GT329 transistors, $B = 570$ MHz; $t_2 = 0.5$; $k = 6$. The difference between the true behavior of the amplifier noise coefficient and that calculated using the empirical formula (3) was no more than 10%.

For the case when $m = 1$, by substituting (3) into (2), and integrating in the frequency band Δf, we can find the average receiver noise temperature values:

$$T_{av} = L_0 T_0 \left[\frac{1}{f_2 - f_1} \ln \frac{f_2}{f_1} + \frac{B^{-k}\,(f_2^{1+k} - f_1^{1+k})}{(1+k)(f_2 - f_1)}\right] + T_{re} \qquad (4)$$

where $T_{re} = T_s + L_0(t_0 + t_2)$. Similarly, when $\begin{cases} m > 0 \\ m \neq 1 \end{cases}$, the following formula holds for the average noise temperature:

$$T_{av} = L_0 T_0 \left[\frac{A^m (f_2^{1-m} - f_1^{1-m})}{(1-m)(f_2 - f_1)} + \frac{B^{-k}(f_2^{1+k} - f_1^{1+k})}{(1+k)(f_2 - f_1)} \right] + T_{re}. \qquad (5)$$

Let us find the limiting IFA frequencies for which the mean square of the temperature fluctuations has the minimal value. To this end, we substitute into (1), T_{av} from (4) or (5), respectively, /84 and find the partial derivatives with respect to f_1 and f_2. After differentiating and some transformations, we have

$m = 1$:

$$\begin{cases} 1 - \dfrac{1{,}5 f_2 \left[1 - \left(\dfrac{f_1}{f_2} \right)^{1+k} \right]}{(f_2 - f_1)(1+k)} + \dfrac{f_0^{1+k}}{f_2^k} \left[\dfrac{1}{f_2} - \dfrac{1{,}5 \ln \dfrac{f_2}{f_1}}{f_2 - f_1} - S \right] = 0, \\ 1 - \dfrac{1{,}5 f_1 \left[\left(\dfrac{f_2}{f_1} \right)^{1+k} - 1 \right]}{(f_2 - f_1)(1+k)} + \dfrac{f_0^{1+k}}{f_1^k} \left[\dfrac{1}{f_1} - \dfrac{1{,}5 \ln \dfrac{f_2}{f_1}}{f_2 - f_1} - S \right] = 0; \end{cases} \qquad (6)$$

$m \neq 1; \quad m > 0$:

$$\begin{cases} 1 - \dfrac{1{,}5 f_2 \left[1 - \left(\dfrac{f_1}{f_2} \right)^{1+k} \right]}{(1+k)(f_2 - f_1)} + \dfrac{f_0^{m+k}}{f_2^k} \left[\dfrac{1}{f_2^m} - \dfrac{1{,}5 (f_2^{1-m} - f_1^{1-m})}{(1-m)(f_2 - f_1)} - S \right] = 0, \\ 1 - \dfrac{1{,}5 f_1 \left[\left(\dfrac{f_2}{f_1} \right)^{1+k} - 1 \right]}{(1+k)(f_2 - f_1)} + \dfrac{f_0^{m+k}}{f_1^k} \left[\dfrac{1}{f_1^m} - \dfrac{1{,}5 (f_2^{1-m} - f_1^{1-m})}{(1-m)(f_2 - f_1)} - S \right] = 0. \end{cases} \qquad (7)$$

Here, the notations are:

$$S = \frac{T_{re}}{2 T_0} \frac{1}{\Delta f^m}; \quad f_0^{m+k} = A^m B^k \neq 0.$$

Thus, after determining experimentally the heterodyne klystron noise parameters and measuring the IFA noise temperature in the passband, we find the values of m, k, F_0. Then, solving the system of equations (6) or (7) for the given receiver noise temperature T_{re}, we determine the limiting IFA frequencies for which the radiometer will have the maximal fluctuation sensitivity.

If we use in the radiometer a mixer of balanced construction and it is known that the heterodyne noise is suppressed by a factor of β, then in (4) and (5) we must substitute $A \cdot \beta^{-1/m}$ in place of

the coefficient A. The calculation of the limiting frequencies was made on a computer. It was found that at the limiting frequencies, the noise temperature increase amounts to 3 dB in comparison with its value at the center of the passband.

Let us return to (2) and (3). After substituting (3) into (2), we have

$$T = L_0 T_0 \left[\left(\frac{A}{f} \right)^m + \left(\frac{f}{B} \right)^k \right] + T_{re}. \qquad (8)$$

This formula (8) describes the superheterodyne receiver noise temperature and includes the noise of the components making up the receiver. However, this same noise temperature can be ascribed to only a single component — the IFA, and all the other components (klystron, mixer, antenna) can be considered noiseless. Then, in order to find the limiting IFA frequencies for which the radiometer will have the minimal value of the fluctuation sensitivity Δt, it is sufficient to find the frequencies corresponding to the points of intersection of the curve plotted using (8) and the horizontal straight line corresponding to increase of the noise temperature by 3 dB relative to its minimal value in the amplifier passband. The required values of A, B, k, m, T_{av} are obtained from experimental study of the klystron, mixer, and IFA used in the radiometer. Thus, for one of the K-45 klystrons and an IFA constructed using GT329 transistors, the following values were obtained: A = 230 MHz; B = 570 MHz; m = 2; k = 6. It is easy to find that the lower amplifier limiting frequency is f_1 = 250 MHz, and the upper is f_2 = 525 MHz for T_{re} = 0. If T_{re} = 1450° K, we find the corresponding values f_1 = 94 MHz, f_2 = 776 MHz. We see from this example that the IFA band is determined by the relationship between the receiver self-noise T_{re} and the noise introduced by the heterodyne and IFA. /85

For minimization of the superheterodyne radiometer fluctuation sensitivity, it is very important to carry out studies of the noise properties of the microwave power generators, which operate as heterodynes, parametric amplifier pumping sources, and so on. The

actual noisy klystron can be represented in the form of two generators: an ideal non-noisy generator at the basic signal frequency, and a noise generator. If, in the superheterodyne radiometer, we use a mixer operating in the wideband regime, i.e., one which along with the signal frequency also receives the mirror frequency, and as the heterodyne we use the subject klystron at the basic frequency, then the noise generator can be broken down into two noise generators with center frequencies $\omega_h - \omega_{if}$ and $\omega_h + \omega_{if}$, respe. ively, where ω_h is the heterodyne frequency and ω_{if} is the intermediate frequency. Such a regime is widely used for shifting the spectrum into the videc frequency band, where its further processing takes place.

In the general case, the noise generated by the klystron generator can be broken down into amplitude-modulated noise (AM noise) plus frequency-modulated noise (FM noise) plus the background noise [4, 5]. It was shown in [6] that the spectral density of FM noise and AM noise as a function of the intermediate frequency magnitude (IF) is the same, beginning with frequencies higher than 20 MHz. Consequently, for IF above 30 MHz, the determination of the magnitude and nature of the behavior of the AM noise plus background noise makes it possible to evaluate the overall klystron noise radiation.

Figure 3 shows a block diagram of a measurement setup which makes it possible to measure the total noise temperature of the system and the AM noise, plus background noise of the heterodyne. We can show that this is so if we consider that the signal spectrum for AM noise:

$$e(t) = C_0\left[\sin\omega_0 t + \frac{M}{2}\sin(\omega_0+\Omega)t + \frac{M}{2}\sin(\omega_0-\Omega)t\right], \qquad (9)$$

and for FM noise and small modulation indices m << 1, which is the case in the klystron, the signal spectrum

$$e(t) = C_0\left[\sin\omega_0 t + \frac{m}{2}\sin(\omega_0+\Omega)t - \frac{m}{2}\sin(\omega_0-\Omega)t\right]. \qquad (10)$$

Thus, the mixer operating in the wideband regime, in which the subject klystron is at the same time both the heterodyne and the noise

signal source, transforms at the IF only the sideband AM noise and the background noise, while the FM noise side frequencies are not transformed at the IF (they cancel one another). The line shown in Figure 3 is used as a matched detector section and permits monitoring the degree of matching of the mixer with the circuit.

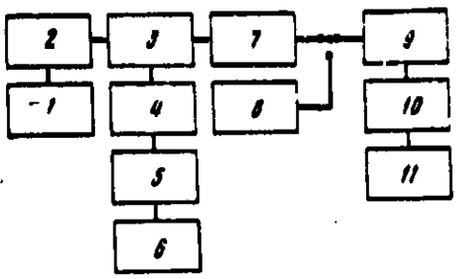


Figure 3. Block diagram of measurement setup:

1 — noise generator; 2 — measuring line; 3 — directional coupler; 4 — stepless HF attenuator; 5 — gate; 6 — test klystron; 7 — frequency converter; 8 — converter output equivalent; 9 — low-noise IF amplifier; 10 — precision IF attenuator; 11 — indicator

/86

The measured mixer relative noise temperature [7] is:

$$t' = \frac{1}{L_0}\left[\frac{N_n(\omega)}{L} + \left(1 - \frac{1}{L}\right)\right] + t_0, \qquad (11)$$

where $N_n(\omega)$ is the spectral density of the AM noise plus background noise of the klystron; L is the noise signal attenuation in the circuit of the subject klystron. In mixers constructed using a Schottky barrier diode with small values of the loss resistance R_s, the quantity t_0 is calculated from the formula:

$$t_0 = t_d\left(1 - \frac{1}{L_0}\right), \qquad (12)$$

where t_d is the relative noise temperature of the Schottky barrier diode (for an ideal diode, $t_d \simeq 0.5$). We see from (11) that the measured quantity consists of three noise components: AM noise plus background noise of the subject klystron, thermal noise in the subject klystron circuit, and the crystal mixer self-noise. By measuring L, L_0, and t_0, we can determine the generator noise radiation spectral density.

The table shows spectral density N_n of the AM noise plus background noise for several klystron types relative to the carrier frequency power level N_c referred to the band $\Delta f = 1$ kHz. Measurements

TABLE

Klystron	6-49				6-49				6-49				6-45			
f, MHz	30	60	100	200	30	60	100	200	30	60	100	200	30	60	100	200
N_n/N_s, dB/kHz	−102	−105	−108	−117	−108	−110	−112	−116	−95	−105	−109	−112	−113	−122	−125	−131
U_r, V	1520				1800				—				1800			
U_{ref}, V	−180				−180				—				−400			
U_f, V	−40				−85				—				−70			
N_s, mW	12				8				4				50			

made at an IF f ≥ 500 MHz showed that the mixer relative noise temperature does not vary with frequency and is equal to

$$t' = \frac{1}{L_0}\left[\frac{1}{L} + \left(1 - \frac{1}{L}\right)\right] + t_0 = \frac{1}{L_0} + t_0. \qquad (13)$$

We can see from the table that the AM noise plus background noise of the klystrons is quite high, and even at frequencies f ∿ 200 MHz N_n $T_0 \sim 3 \cdot 10^4$ °K. When measuring N_n (ω) of klystrons as a function of frequency, it is necessary to know the mixer conversion losses. This quantity is found as

$$L_0 = \frac{N_1}{N_2 - (t' - 1)}, \qquad (14)$$

where N_1 is the noise generator spectral density at the mixer input; N_2 is the noise radiation spectral density at the mixer output terminals. The quantity N_2 was measured by replacing the HF noise generator plus mixer system by an equivalent IF noise generator with known and controllable noise temperature. In this case, the IF noise generator had impedance equal to the mixer output impedance.

We investigated the dependence of klystron noise radiation spectral density on the magnitude of the voltage U_f on the focusing electrode. Figure 4 shows the results of measurement of the noise /87

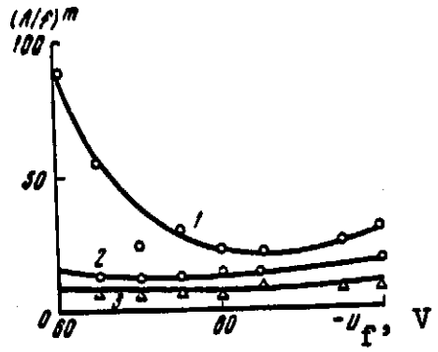


Figure 4. Experimental dependence of type K-49 klystron noise radiation spectral density on focusing electrode voltage:

U_r = 1800 V; U_{ref} = -170 V;
1 — IF 30 MHz; 2 — 60 MHz;
3 — 100 MHz

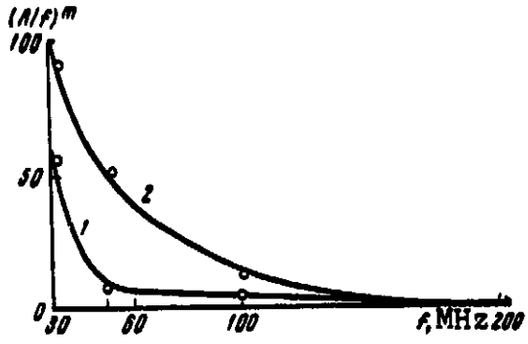


Figure 5. Nature of experimental dependence of redundant relative noise temperature of frequency transformer:

1 — klystron K-45; 2 — K-49

spectral density of a type K-49 klystron, referred to the mixer output, for different IF and constant L_0. We see that the noise generated by the klystron depends strongly on U_f (the voltage U_r on the resonator and U_{ref} on the reflector were held constant, and during the measurement time the mixer crystal current was held constant) at frequencies below f < 30 - 50 MHz, and depends weakly on U_f at frequencies f > 50 MHz. The klystron noise is minimal for U_f corresponding to the maximal generated signal power. Moreover, during the measurement, we observed the hysteresis phenomenon, i.e., the minimum of the relative noise temperature shifts in one direction or another, depending on how U_f changes (from -60 to -100 V or from -100 to -60 V). Therefore, an intermediate frequency f_{IF} > 50 MHz should be selected in mm- and cm-band superheterodyne receivers.

Analysis of the experimental results obtained (Figure 5) for t' made it possible to obtain the empirical formula (2), which was used in calculating the limiting IFA passband frequencies. After the limiting frequencies are determined, it is necessary to construct

the amplifier frequency characteristic in some fashion or other (for example, with the aid of filters).

REFERENCES

1. Basharinov, A. Ye., A. S. Gurvich and S. G. Yegorov. Doklady AN SSSR, Vol. 188, 1969, p. 1273.

2. Fischetti, T. L., D. L. Lind and O. G. Smith. Astronautics and Aeronautics, Vol. 9, 1971.

3. Ekspress-informatsiya, Seriya "Radiolokatsiya, televideniye, radiosvyaz'" (Express Information, Series Radar, Televison, Radio). Vol. 6, 1971, p. 27.

4. Bosch, B. G. and W. A. Gambling. Brit. I.R.E., Vol. 22, 1962, p. 389.

5. Bosch, B. G. and W. A. Gambling. Brit. I.R.E., Vol. 21, 1961, p. 503.

6. Rao, B. V. and W. A. Gambling. Radio Elect. Eng., Vol. 35, 1968, p. 165.

7. Nikulina, L. N., Yu. N. Nemlikher and I. A. Strukov. In the collection: Poluprovodnikovye pribory i ikh primeneniye (Semiconductor Devices and their Application), Vol. 27. Moscow, 1972.

STUDY OF SCHOTTKY BARRIER DIODE FREQUENCY CONVERTER IN THE SHORT MILLIMETER WAVELENGTH BAND

V. F. Kolomeytsev, Yu. Yu. Kulikov, A. M. Kupriyanov, I. A. Strukov, L. I. Fedoseyev, Yu. B. Khapin and V. S. Yetkin

ABSTRACT. We present a technique for and results of measurement of Schottky barrier diode basic parameter[illegible]t 1.8 - 2 mm wavelengths, and also comparison of these data with the results obtained on point-contact silicon diodes. We discuss the possibilities for the use of Schottky barrier diodes in frequency converters for the short millimeter wavelength band. According to the measurement results, the relative noise temperature of the Schottky barrier diode is approximately 1/2, and the diode ideality parameter is 1.2. The frequency converter relative noise temperature is close to 1. The measured conversion losses in the radiometric regime were less than 13 dB at λ = 1.88 mm.

The absence of short millimeter band UHF leads to the necessity /88 for constructing radiometers for this band using the scheme: modulator-frequency converter-IFA-detector. In this case, the radiometer reliability and stability are determined by the reliability and stability of the frequency converter, in which point-contact diodes are used [1, 2]. The latter have poor stability, reproducibility, and reliability, which limits application of short millimeter

band radiometers for the solution of a whole series of problems of radiophysics, radioastronomy, and thermal radar.

Recently, considerable attention has been devoted, both in the USSR and abroad, to study of majority-carrier diodes (Schottky barrier diodes fabricated using planar-epitaxial technology). Just as the point-contact diodes, the Schottky barrier diodes[1] utilize the rectifying properties of a metal semiconductor contact. The use of planar-epitaxial technology has made it possible to create a "honeycomb" structure with reproducible parameters which do not change with time. In the Schottky barrier diodes, there is no minority carrier accumulation effect and, consequently, there is no diffusional capacitance. Thanks to this property, these diodes are fast-acting varistors, and may find application in microwave switching devices and microwave frequency down-converters. The use of a Schottky barrier diode in a 5-mm frequency converter was reported in [3]. Conversion losses of 5 - 6 dB were obtained. As far as we know, no experiments have been conducted on use of Schottky barrier diodes in frequency converters at shorter wavelengths.

The purpose of the present article is to investigate the possibility of using Schottky barrier diodes in frequency converters in the short millimeter segment of the spectrum, and to conduct an experiment using a Schottky barrier diode and a point-contact diode. The experimental conditions were the same in both cases. Diodes without holders were used in both cases. The studies were conducted in the same chamber. The dimensions of the pins for the p/p structures and the needle were also the same. No effort was made to optimize the characteristics of the frequency converter. A gold-GaAs contact of n-type was used as the rectifying contact of the Schottky barrier diode, a tungsten-silicon contact was used as the point-contact diode.

---

[1] Hereafter, we shall use for brevity the term "Schottky barrier diode" to mean "Schottky barrier diode fabricated using planar-epitaxial technology".

The frequency converter transfer, noise, and impedance characteristics are defined by the Volt-Ampere characteristic. Figure 1 shows the V-A characteristics of two diodes: Si point-contact and GaAs Schottky barrier and their equivalent circuit. The V-A characteristics of both diode types are described analytically by the expression:

$$I = I_s\left[\exp\left(\frac{qU}{nkT}\right) - 1\right], \qquad (1)$$

where n is a dimensionless coefficient; I_s is the saturation current; q is the electron charge; k is the Boltzmann constant; T is the absolute temperature. We rewrite (1) in the form:

$$I = I_s\left[\exp\left(\alpha U\right) - 1\right]. \qquad (2)$$

/89

The quantity α is determined from the slope of the semilog V-A characteristic:

$$\alpha = 2{,}303\frac{\Delta(\lg I)}{\Delta U}. \qquad (3)$$

In the ideal diode case, n = 1 and

$$\alpha = \frac{d\ln I}{dU} = q/kT = 38{,}6b^{-1} \quad \text{for } T = 300^\circ\text{K}.$$

Figure 2 shows the diode V-A characteristics on a semilog scale. For real diodes, the V-A characteristic slope is α = 30 V^{-1} for the

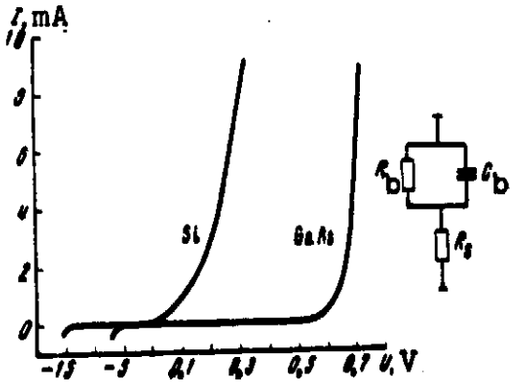


Figure 1. V-A characteristics of Si point-contact and GaAs Schottky barrier diodes and their equivalent circuit

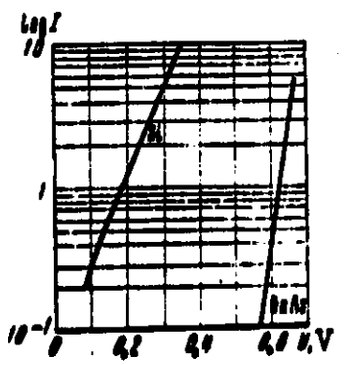


Figure 2. Semilog V-A characteristics of Si point-contact and GaAs Schottky barrier diodes

Schottky barrier diode, and $\alpha = 12\ V^{-1}$ for the point-contact diode. Thus, the parameter n for the corresponding diodes is: $n = 1.28$ for the Schottky barrier diode, $n = 3.2$ for the point-contact diode. The saturation currents for the corresponding diodes are determined by the points of intersection of the V-A characteristics with the current axis. For the GaAs Schottky barrier diode, $I_s = 4.5 \times 10^{-4}$ A; for the point-contact diode, $I_s = 10^{-4}$ A. The diode series resistance R_s can also be determined from the V-A characteristics: for the Schottky barrier diode, $R_s = 15$ Ohm, for the point-contact diode, $R_s = 20$ Ohm.

An important characteristic of frequency down-converter diode quality is the noise ratio N_R, i.e., the ratio of the diode nominal noise power to the noise power of a resistor whose magnitude is equal to $R_b + R_s$. It is not difficult to show that for the Schottky barrier diode, the resulting relation for N_R has the form:

$$N_R = \frac{(n/2)R_b + R_s}{R_b + R_s}. \qquad (4)$$

For small currents through the diode, when $R_b >> R_s$, we can obtain $N_R = 0.5$ if $n = 1$. With increase of the current through the diode, N_R slowly increases because of R_s noise.

In order to exclude the influence of heterodyne noise on the frequency converter noise characteristics, the intermediate frequency is selected in the 3-cm band. The block diagram of the setup for measuring the noise ratio and noise temperature of diodes in this band is shown in Figure 3. The measurement error amounted to 1° K. The noise temperature measurement results of the corresponding diodes are shown in Figure 4. The GaAs Schottky barrier diode has significantly lower noise temperature in comparison with the Si point-contact diode. Another disadvantage of point-contact diodes is the considerable noise temperature change with "overheating" of the points (shaded region). Curves of the noise ratio of the

/91

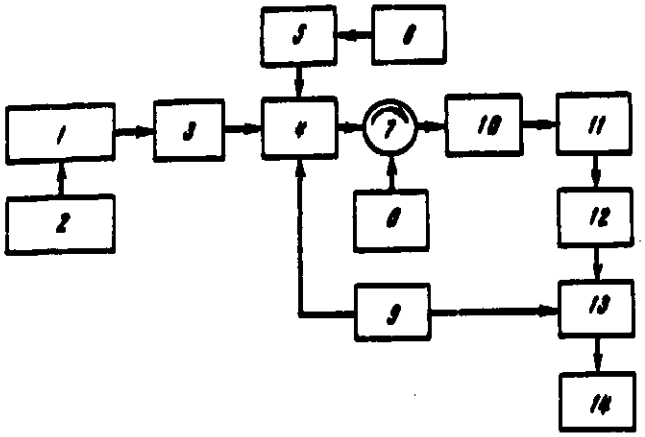


Figure 3. Experimental setup for determining diode noise ratio and noise temperature

1 — waveguide chamber; 2 — external constant bias; 3 — measuring line; 4 — modulator; 5 — precision attenuator; 6 — noise generator; 7 — circulator; 8 — random signal generator; 9 — reference voltage generator; 10 — tunnel amplifier; 11 — detector; 12 — LFA; 13 — superheterodyne demodulator; 14 — recorder

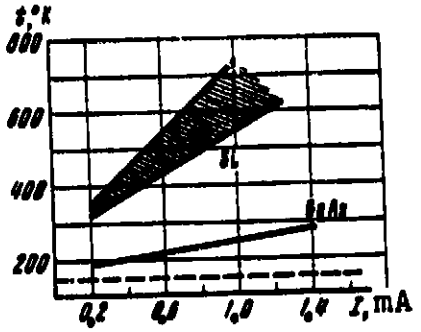


Figure 4. Noise temperature of Si and GaAs diodes versus current through the diodes

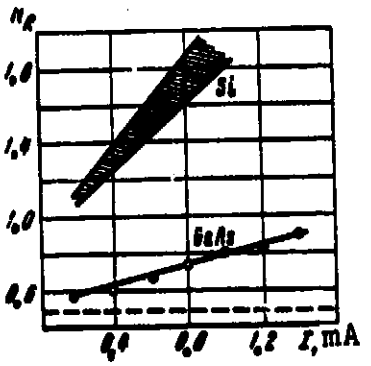


Figure 5. Noise ratio of Si and GaAs diodes versus current through the diodes:

1 — theory; 2 — experiment

corresponding diodes as functions of the current through the diodes are shown in Figure 5. The theoretical and experimental noise ratios of the Schottky barrier diode agree.

The basic parameters characterizing the frequency converter are the conversion loss L and relative noise temperature t. The technique for measuring the relative noise temperature does not differ from the technique for measuring diode noise ratio. The relative noise temperature was found to be equal to t = 1.2 and t = 2 - 2.5 for the Schottky barrier and point-contact diodes.

The conversion losses were measured using the scheme shown in Figure 6. The input signal was the noise signal of a GSh-8 4-mm tube. Since the spectral density of its radiation decreases sharply in the 2-mm part of the spectrum, the tube was calibrated at the wavelength $\lambda = 1.88$ mm with the aid of a detector receiver. A blackbody in boiling nitrogen was used as the primary reference. The frequency converter conversion losses measured in this fashion in the two-channel regime at $\lambda = 1.88$ mm were 13 and 16 dB for the Schottky barrier and point-contact diodes, respectively.

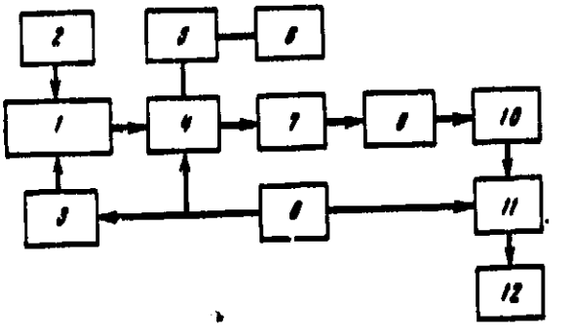


Figure 6. Experimental setup for determining conversion losses:
1 — frequency converter; 2 — heterodyne; 3 — GSh-8; 4 — modulator; 5 — precision attenuator; 6 — noise generator; 7 — tunnel amplifier; 8 — reference voltage generator; 9 — detector; 10 — LFA; 11 — superheterodyne detector; 12 — recorder

REFERENCES

1. Bauer, B. J. A Low Noise Figure 94 GHz GaAs Mixer Diode. The Microwave Journal, Vol. 9, No. 10, 1966, p. 84.

2. Kislyakov, A. G., Yu. V. Lebskiy and A. I. Naumov. Izv. VUZov, seriya radiofizika, Vol. 11, No. 12, 1968, p. 1791.

3. Young, D. T. and J. C. Irvin. Proc. IEEE, Vol. 53, No. 12, 1965, p. 2356.

INFLUENCE OF PHASE SHIFTER ON FREQUENCY DIVIDER CHARACTERISTICS

Ya. E. Veyber

ABSTRACT. We study the influence of a phase shifter in the feedback loop on circuit characteristics for the example of a frequency divider with converter and amplifier. It is shown that use of the phase shifter is not advisable in most cases. We find the external signal amplitude at which the synchronism band reaches a maximum for even division ratios $n \geq 4$.

INTRODUCTION

In the last 10 - 15 years, frequency dividers (FD) of the regenerative type with multi-order frequency conversion have been developed and studied [1, 2, 4, 7 - 9], which have a simple circuit (absence of frequency multiplier), wide synchronism band, and ease of regulation. This variety of regenerative FD includes the frequency dividers with converter and amplifier (DCA) [1 - 5], the push-pull FD [7, 8], and the push-pull FD with reactive feedback [9, 10]. In these FD there are no asynchronous oscillations outside the synchronism band [11] because of the smallness of the so-called asynchronous current component (I_{01}). Some researchers [3, 6, 12] have suggested the introduction of a phase shifter into the feedback loop in order to improve the FD parameters significantly. According to these investigators, the introduction of "optimal" phase shift maximizes the mixer transmission coefficient, which increases the synchronism band and stability of the division process. According to Andreyev and Tseitlin [4, 13], FD circuits without phase shift

/91 /92

in the feedback loop (i.e., without a phase shifter) are usually used, as a result of which maximal synchronism band and maximal output voltage phase stability are obtained. However, there is a phase shifter in the circuit proposed in [4]. Reference [14] also indicates the absence of any marked improvement in the synchronism band in the FD with small value of I_{01} with the introduction of complex feedback.

In the present article, we investigate the influence of a phase shifter on FD characteristics and present recommendations relative to its use in FD circuits. We study the influence of the phase shifter for the example of a DCA circuit whose transistor version is shown in Figure 1. The frequency converter (FC) is a bridge consisting of the diodes D_1 - D_4, to which there is applied the input signal e and the feedback voltage u from the cascade amplifier consisting of the transistors T_1 and T_2.

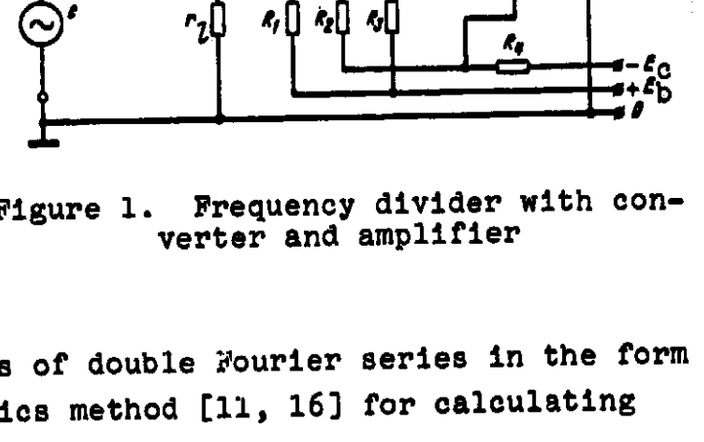


Figure 1. Frequency divider with converter and amplifier

We make the study using the method of slowing varying amplitudes, we use the symbolic method of [15] for writing the equations, and we use the apparatus of double Fourier series in the form of the modulation characteristics method [11, 16] for calculating the current spectrum at the FC output.

FREQUENCY DIVIDER EQUATIONS

We shall make several assumptions which simplify analysis of the circuit (Figure 1).

1. We consider that the amplifier operates in the linear regime and there is no collector current reaction (as, for example, in the cascade amplifier).

2. The inductance of the feedback loop winding has no influence on FD bridge operation. The admissibility of this assumption is determined by the small magnitude of the modulus of the feedback coefficient k for transistor circuits ($k \sim 0, 1$).

3. The stray inductance of the circuit coil is equal to zero. For the frequency range 0.1 - 30 MHz in FD circuits, coils are used with magnetoelectrically armored cores, having small stray inductance even in comparison with the feedback winding inductance.

4. We assume that in the amplifier, transistors are used with limiting current amplification frequency f_α, exceeding by 5 - 10 times the subharmonic frequency f_1, so that we can ignore the inertial properties of the current carriers in the transistor.

Let us formulate the equations of the DCA. Assume that the external synchornizing signal $e = E \cos \omega t$, and the feedback voltage:

$$u = U(t) \cos [\omega_1 t + \varphi(t)]. \tag{1}$$

act on the FC bridge. The complex feedback voltage amplitude: /93

$$\dot{U} = Ue^{j\varphi} \tag{2}$$

is connected with the complex amplitude of the current through the loop I_1 and the symbolic control impedance Z_y (p) by Ohm's law:

$$\dot{U} = \dot{I}_1 Z_y(p). \tag{3}$$

The magnitude of the symbolic impedance for the circuit with a single parallel loop is equal to:

$$Z_y(p) = \frac{R_{0e}}{1 + pT + p^2 T^2} k = \frac{R_{0e} k}{1 + pT + p^2 T^2} e^{j\varphi_k}, \tag{4}$$

where R_{0e} is the equivalent active resistance of the loop with time constant T at the natural frequency ω_0; $p = d/dt$ is the differential

operator; $\xi = (\omega/n - \omega_0)\,T$ is the generalized loop detuning; $k = k_c^{j\phi k} = U/U_k$ is the feedback coefficient with modulus k and ϕ_k; U_k is the complex amplitude of the loop voltage. Substituting the expression presented below for the complex current amplitude (12) and also (2) and (4) into (3), separating the real and imaginary parts, we obtain a system of two abbreviated equations describing the processes in the FD in the synchronism regime:

$$\begin{aligned} TpU &= U(N\gamma_c \cos n\varphi \cdot \cos\varphi_k + N\gamma_s \sin n\varphi \cdot \sin\varphi_k - 1), \\ Tp\varphi &= -(N\gamma_s \sin n\varphi \cdot \cos\varphi_k - N\gamma_c \cos n\varphi \cdot \sin\varphi_k + \xi), \end{aligned} \tag{5}$$

where $N = S_e R_{0e} k$ is the regeneration coefficient; S_e is the equivalent slope of the characteristics of the FC-amplifier circuit, γ_c,s are the in-phase and quadrature with respect to feedback voltage subharmonic current expansion coefficients (EC).

RESULTS OF FC OPERATION ANALYSIS

Using the double Fourier series apparatus [11, 16], we determine the instantaneous value of I_1 and the complex amplitude of $i_{\omega/n}$ of the current components (with frequency $\omega_1 = \omega/n$) falling in the LC loop passband for the synchronism regime:

$$i_{\omega/n} = R_l S_t \left[I_{1,n-1} \cos\left(\frac{\omega}{n} t + \varphi - n\varphi\right) + I_{1,n+1} \cos\left(\frac{\omega}{n} t + \varphi + n\varphi\right) \right], \tag{6}$$

$$I_1 = (I_c \cos n\varphi - jI_s \sin n\varphi) e^{j\varphi}, \tag{7}$$

where

$$I_{c,s} = R_l S_t (I_{1,n-1} \pm I_{1,n+1}). \tag{8}$$

/95

Here, S_t is the slope of the transistor (amplifier) characteristic; n is the division ratio; R_l is the equivalent resistance of the FC load, determined by the amplifier input resistance R_{in} and the converter load resistance r_l (see Figure 1). In components of the form I_{1l}, the first subscript denotes the number of the current harmonic

through the FC load with expansion into a Fourier series in the external signal frequency ω, and the second subscript is the harmonic number with expansion of I_1 into a Fourier series in the synchronous frequency $\omega_1 \equiv \omega/n$.

We shall approximate the Volt-Ampere characteristic of the FC bridge diodes by a polygon with cutoff ($E_o > 0$ is the cutoff voltage, r_o is the resistance of the open diode) in order to take into account the influence of the cutoff voltage E_o on the FD parameters, since for the small FC loads characteristic for transistor circuits, the voltages applied to the FC are commensurate with the cutoff voltage E_o. In this case, the amplitude of the first current harmonic I_1 (u) with frequency ω (modulation characteristic) is equal to /96

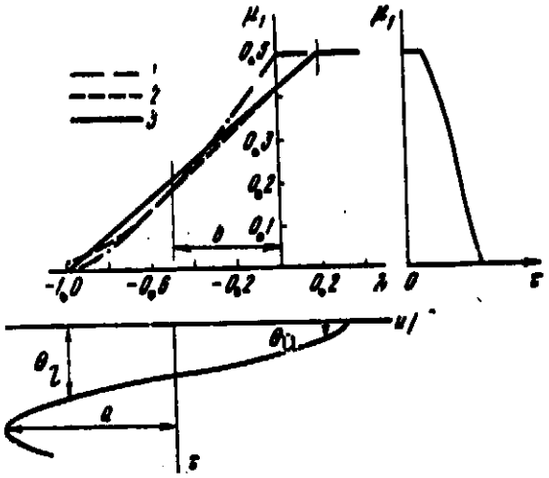


Figure 2. Approximation of modulation characteristic by polygon:

The position of the operating point on the characteristic is determined by the signal amplitude E; for large feedback voltage amplitudes U, two cutoff points appear; 1 — theoretical data; 2 — experimental data; 3 — approximation

$$\frac{I_1(u)}{S_{dyn}E} = \begin{cases} \gamma_1(0), & u - 2E_0 \leqslant 0, \\ 0{,}5, & u - 2E_0 > 0, \end{cases}$$

where γ_1 (0) is the EC for the first current harmonic [16], and the diode bridge dynamic slope S_{dyn} and current cutoff angle θ are defined by the equalities

$$S_{dyn} = \frac{1}{r_0 + R_1 + R_2}; \quad \cos\theta = -\frac{u - 2E_0}{E}. \tag{9}$$

In order to facilitate the calculation of the components I_{1i}, we approximate the modulation characteristic I_1 (u) by a polygon, considering, in so doing, the experimental data (Figure 2). As a

result of the calculations, we obtain:

$$I_{1l} = S_{dyn}\, Esa\, [\gamma_l\ (\theta_l) - \gamma\ \ (\theta_u)] = 0.417\ S_{dyn}\ U \cdot$$

$$0.417\ S_{dyn}\ U\ [\gamma_l\ (\theta_l) - \gamma_l\ (\theta_u)], \qquad (10)$$

where $\gamma_l\ (\theta)$ is the EC for the l^{th} current harmonic [16]; s = 0.5/1.2 = 0.417 is the slope of the approximating straight line; a = U/E and b = 2 E_0/E are the relative values of the feedback and bias voltage amplitude; θ_l and θ_u are the lower and upper current cutoff angles, defined by the equalities:

$$\cos\theta_l = \frac{b-1}{a},\ \cos\theta_u = \frac{b+0.2}{a} = \cos\theta_l + \frac{1.2}{a}. \qquad (11)$$

With account for (10), the expressions (7) and (8) take the form

$$I_1 = S_e U\,(\gamma_c \cos n\varphi - j\gamma_s \sin n\varphi)\, e^{j\varphi}, \qquad (12)$$
$$I_{1\ldots} = S_e U \gamma_{c,s}$$

where

$$\gamma_{c,s} = [\gamma_{n-1}(\theta_l) - \gamma_{n-1}(\theta_u)] \pm [\gamma_{n+1}(\theta_l) - \gamma_{n+1}(\theta_u)], \qquad (13)$$
$$S_e = 0{,}417 R_l S_t S_{dyn}.$$

The graphs of the EC γ (a) for various values of the bias b and division ratio n = 2, 4, 5, constructed using (13) and (11), are shown in Figure 3. We see that the bias b (or amplitude of the external signal E) has marked influence on the nature of the EC curves, which confirms the necessity for considering the cutoff E_o.

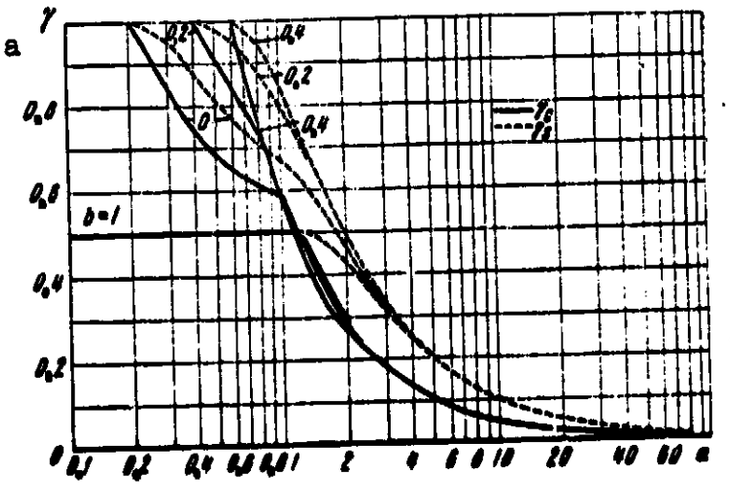


Figure 3. Expansion coefficients γ versus signal amplitude a for various biases b for division ratio:

a — n = 2; b — n = 4; c — n = 5

(Figure continued on following page)

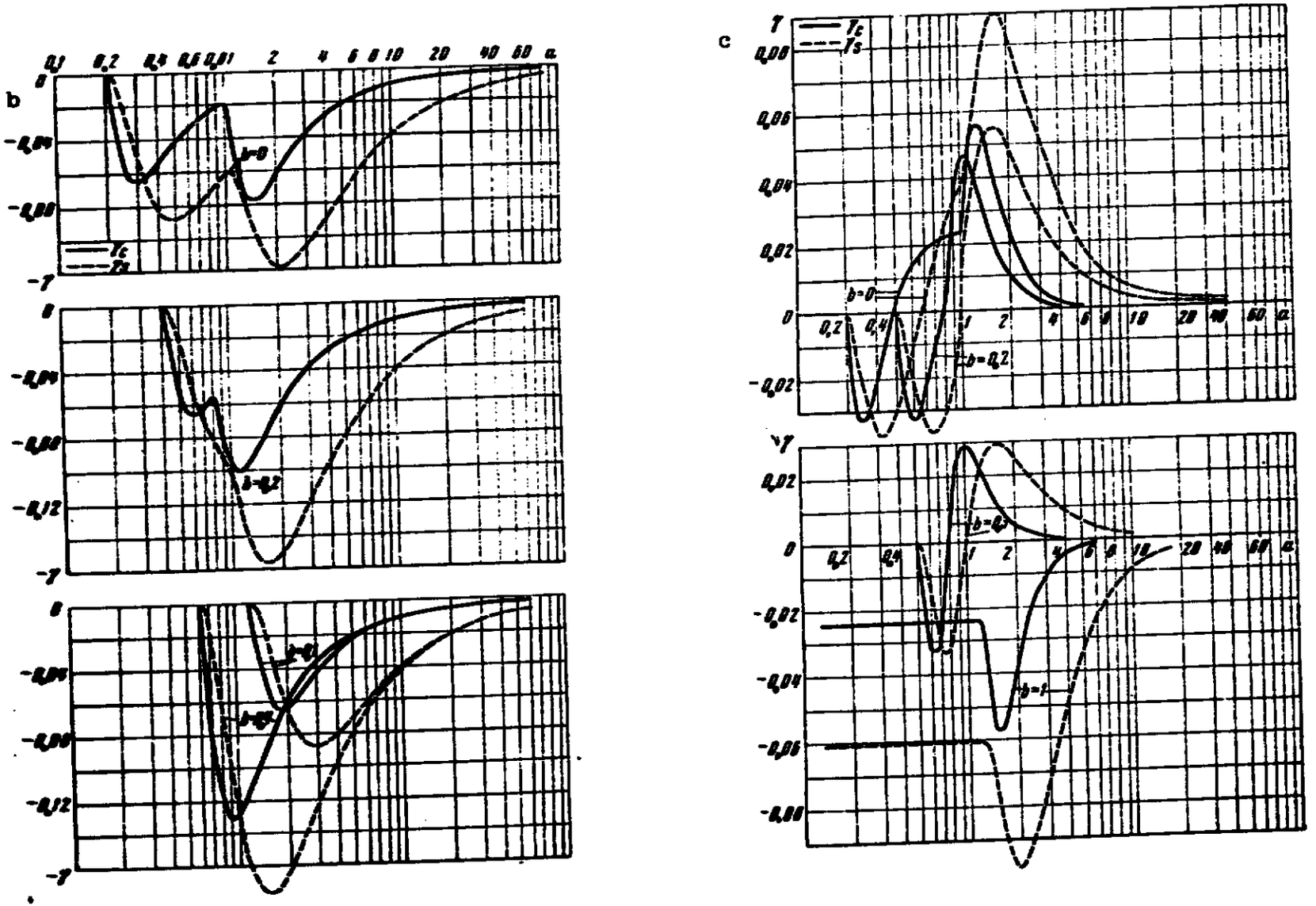


Figure 3. (continued)

Let us examine the influence of phase shifts in the feedback loop on FC operation. According to (6), the amplitude of the subharmonic current in the FC load depends on the feedback voltage ϕ, and the subharmonic phase ψ * does not coincide in the general case with the phase ϕ. It may be shown that, in the region of stable values of the amplitudes a, the signs of the components $I_{1,n-1}$ and $I_{1,n+1}$ are opposite. Therefore, according to (6), the maximal and minimal subharmonic current amplitudes will be obtained for phases equal to

$$\varphi_{I\max} = \pm\frac{\pi}{2n}, \quad \psi_{I\max} = \mp\frac{\pi}{2n}(n-1), \quad \varphi_{I\min} = \psi_{I\min} = 0. \tag{14}$$

The phase angle α of the first loop current harmonic (relative to u), /97 corresponding to the current extrema, can be found on the basis of the equation of FD phase balance, formulated under the assumption of absence of phase shifts in the amplifier (Figure 4):

$$\psi + \alpha + \Phi = \varphi, \tag{15}$$

where $\Phi = \phi_k$ — is the phase shift introduced by the phase shifter. From (14) and (15), we obtain the following values of the angles α:

$$\alpha_{I\max} = \begin{cases} \pm\pi/2, & \Phi = 0, \\ 0, & \Phi = \pm\pi/2; \end{cases}$$
$$\alpha_{I\min} = 0, \ \Phi = 0.$$

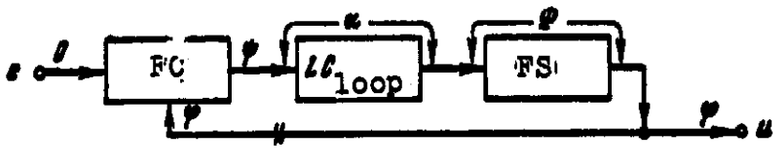


Figure 4. Block diagram of DCA for phase deviations:

FC — frequency converter; FS — phase shifter

Consequently, the largest amplitude of the current components with frequency ω/n at the FC output is obtained with feedback voltage phase $\phi = \pm \pi/2n$, or with load (loop) phase angle $\alpha = \pm \pi/2$. An angle close to this value can be achieved only at the edge of the

* The phase is indicated relative to the phase of the signal e.

synchronism band. When the loop is tuned to resonance with the subharmonic frequency (with the feedback loop open), the current amplitude will be minimal. To obtain a maximum subharmonic current with resonant tuning of the oscillatory system, we can use a coupled system of loops, or a phase shifter which introduces the additional phase shift $\Phi = \pm \pi/2$. We emphasize that the indicated phase shift magnitude is optimal only in the sense of obtaining maximum subharmonic current at the FC output, but the characteristics of the FD as a whole may not be improved in this case.

STATIONARY REGIME CHARACTERISTICS

The FD behavior in the stationary regime is described by the equations:

$$\gamma_c = \frac{1}{N\cos n\varphi}, \qquad \xi = -N\gamma_s \sin n\varphi, \qquad \Phi = 0, \tag{16}$$

$$\gamma_s = \pm\frac{1}{N\sin n\varphi}, \qquad \xi = \pm N\gamma_c \cos n\varphi, \qquad \Phi = \pm\pi/2, \tag{17}$$

obtained from (5) by the substitutions $pU = p\phi = 0$ and $\phi_k = \Phi = 0$, $\pi/2$[1]. We obtain all the basic stationary regime characteristics from (16) and (17).

It is well known that for $n > 2$ in the FD without asynchronous component I_{01}, subharmonics are not excited from the equilibrium state [11]; therefore, in place of the threshold with respect to the input signal E, it is more convenient to use the concept of <u>threshold regeneration coefficient</u> N_{th} (suitable for any n), to which there corresponds the minimal controlling resistance magnitude, when the conditions of existence of the stationary regime in the absence of detuning, $\xi = 0$, are still satisfied:

$$N_{th} = \begin{cases} 1/\gamma_{s\max}, & \Phi = 0, \\ 1/\gamma_{s\max}, & \Phi = \pm\pi/2. \end{cases} \tag{18}$$

---

[1] The resonance and phase characteristics are symmetric only for the phase shift $\Phi = 0, \pi/2$.

Here $\gamma_{c,s\ max}$ denotes the maximum absolute values of the functions γ_c and γ_s. It is obvious that synchronism is possible only for $N > N_{th}$. Figure 5 shows the dependence of N_{th} on external signal amplitude E, plotted from (18) and the EC curves; the dash-dot line corresponds to infinite threshold growth for $b \geq 1$, when the FC does not operate. The curves show rapid increase of N_{th} with increase of n, the presence of a threshold with respect to external input, and the absence of a ceiling for the DCA circuit. For n = 2, the presence or absence of the phase shifter in the circuit does not alter the FD threshold properties, for n > 2 introduction of the "optimal" phase shifter into the circuit reduces the threshold magnitude in the best case by a factor of 1.5 - 2.

/98

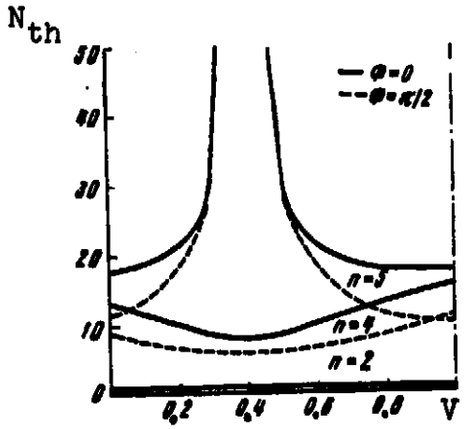


Figure 5. Threshold regeneration coefficient N_{th} versus external signal amplitude E

We note the presence for n > 2 of the function N_{th} (V) for even division ratios n and sharp increase of the threshold for odd n in the vicinity of the point b = 0.4 (or $E = 5\ E_0$), corresponding to the midpoint of the modulation characteristic. This means that the synchronizing signal amplitude $E = 5E_0$ is optimal for even division ratios $n \geq 4$, since in this case the synchronism band reaches a minimum; for odd ratios $n \geq 3$ in the region of E values close to $5\ E_0$, subharmonic generation will terminate.

We obtain the <u>resonance</u> a (ξ) <u>and phase</u> nϕ (ξ) <u>frequency characteristics</u>(1) from solution of the stationary regime equations (16)

(1) We shall consider that the amplitudes and phases of the voltages u_{out} and u (see Figure 1) coincide, although in reality, for $\phi \approx \pi/2$, the phases of these voltages differ by $\pi/2$.

and (17). In order to calculate the characteristics for $\Phi = 0$ using the selected regeneration coefficient $N > N_{th}$, specifying values of $\cos n\phi$ and determining $\sin n\phi$, we find from the first equation (16) the quantity γ_c. From the $\gamma_{c,s}$ (a) graphs, we find the self-maintained oscillation amplitude ξ and the EC γ_s, and from the second equation (16) we find the value of the detuning ξ. We solve (17) similarly. The constructed frequency characteristics for the bias $b = 0.2$ and $n = 2, 4, 5$ are shown in Figure 6[1]. For the division ratio $n = 2$, the frequency characteristics are reminiscent of the characteristics of an oscillatory loop. With increase of n and N for $\Phi = 0$, the DCA frequency characteristics improve (i.e., the resonance curves become flatter and flatter, and the phase curves become prattically linear), while for $\Phi = \pi/2$ the characteristics deteriorate; the resonance characteristics become sharper and sharper, the length of the phase curve linear segment decreases, and the slope of the characteristic curve and the phase rate of advance increase. With $n \geq 4$ for the case $\Phi = 0$, and with $n \geq 5$ for the case $\Phi = \pi/2$, the frequency characteristics become multivalued because of the oscillatory nature of the EC; excitation of large-amplitude oscillations becomes most likely, although in practice we can observe excitation of small-amplitude oscillations. Transition from one branch of the characteristic to the other may be accomplished by jump of the phase $n\phi$ by $\pm\pi$ (Figure 6c), which is undesirable.

Analysis of the frequency characteristics showed significant deterioration with presence of the "optimal" phase shifter in the circuit. Actually, in view of the peaked nature of the resonance curves (for $\Phi = \pi/2$), with even small detuning of the loop relative to the input signal frequency (for example, aging of the oscillatory loop components, operation over a range of temperatures, signal frequency drift, and so on), marked reduction of the output voltage amplitude is possible, which may lead to disruption of the operation

---

[1] The characteristics of the DCA with phase shifter are denoted by the subscript "$\Phi = -90°$".

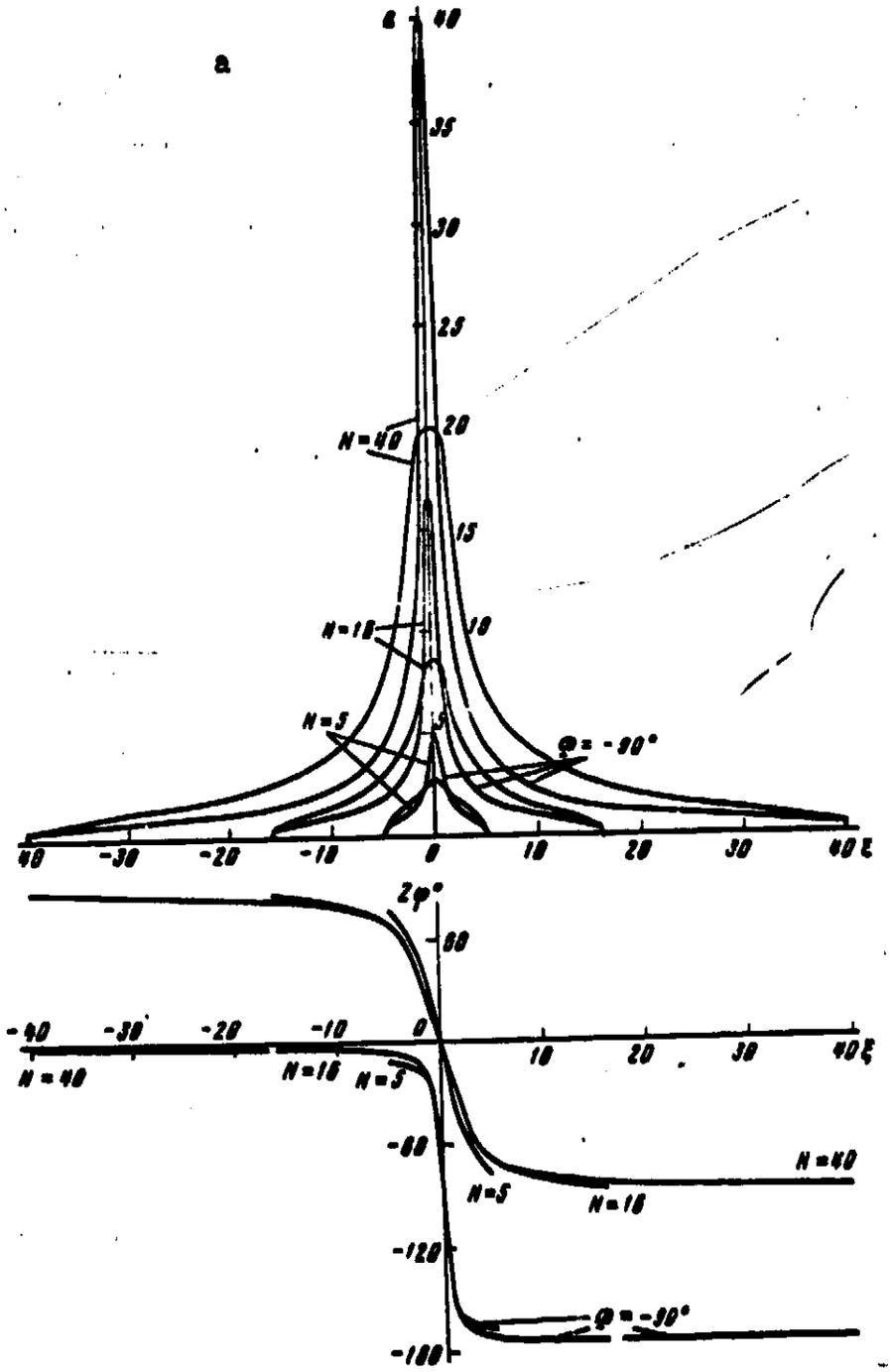


Figure 6. Frequency cha.acteristics of DCA for bias b = 0.2:
a — n = 2

(Figure continued on following page)

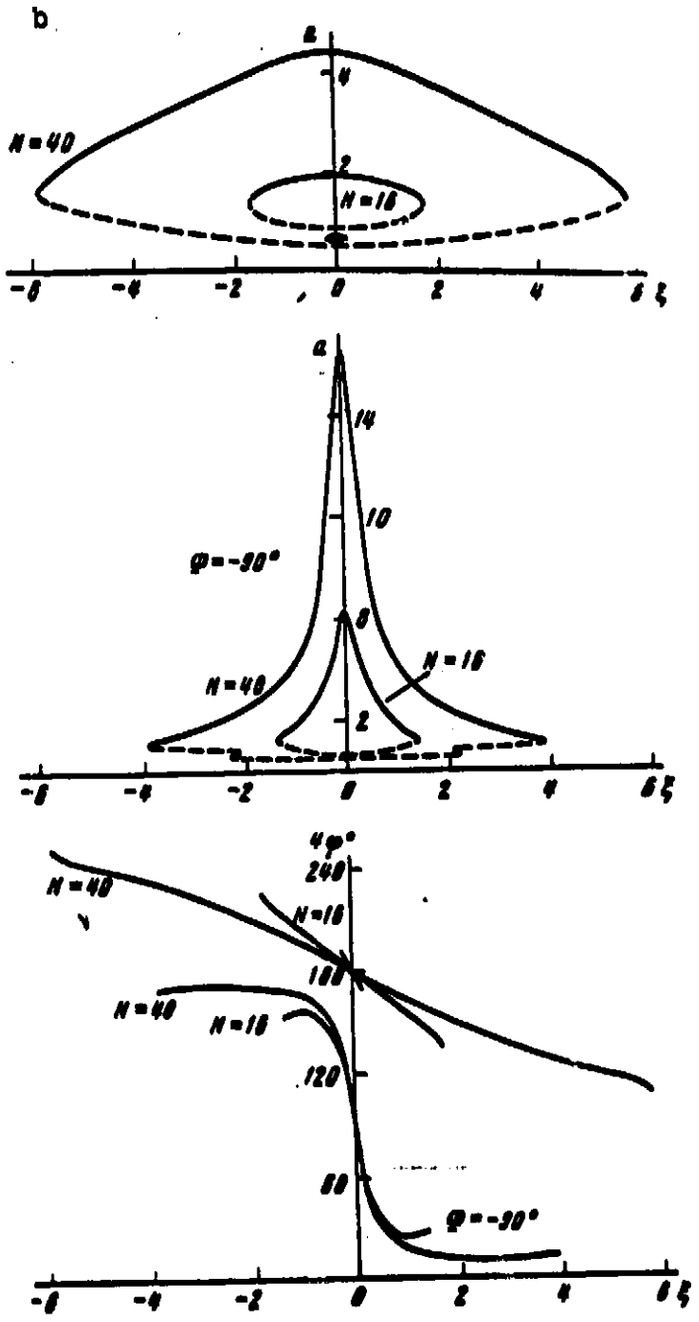


Figure 6. (continued):
b — n = 4

(Figure concluded on following page)

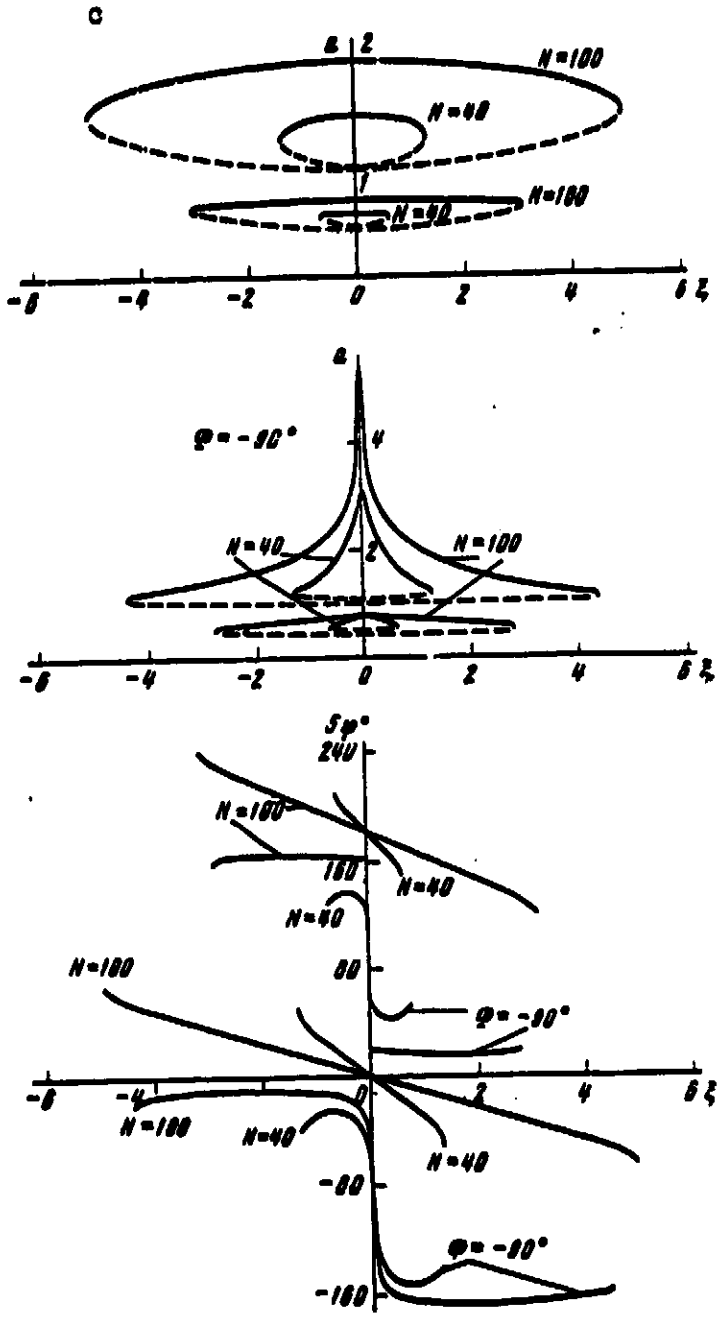


Figure 6. (concluded):
c — n = 5

of the subsequent devices and practical reduction of the operating frequency band, although the synchronism band does not change. On the other hand, for $\Phi = \pi/2$ for any n, the phase characteristics remain nonlinear and the phase advance rate increases significantly (for example, for $n = 5$, $N = 40$, $b = 0.2$, the phase advance rate increases by 18 times), which leads to a corresponding increase of the phase instability. In addition, it is not possible to regulate the degree of resonance curve narrowing. /99

The synchronism band (i.e., the region of maximal possible detunings within the limits of which the synchronism regime is maintained, if it was previously established) can be determined on the basis of the stationary regime stability conditions, or directly from the resonance characteristic, since stability transition takes place at the boundaries of the synchronism band. It is convenient to express the synchronism band Π_c in units of generalized detuning ξ_c by the known equality:

$$\xi_0 = \frac{\Pi_0}{f_0} Q,$$

where f_0 and Q are the natural frequency and equivalent quality factor of the FD loop. Figure 7 * shows curves of DCA synchronism band, expressed in units of detuning ξ_c, versus regeneration coefficient N for division ratios $n = 2, 4, 5$. The relation ξ_c (N) for the case $n = 2$ is given by the formula:

$$\xi_0 = \sqrt{N^2 - 1}; \quad 0 \leqslant b < 1; \quad \Phi = 0, \pi/2, \tag{19}$$

obtained upon substituting into (16) or (17) the limiting EC values $\gamma_c = \gamma_s = 1$. The graphs show clearly that upon connection of the "optimal" phase shifter into the FD circuit for the range of regeneration values $N > (1.5 - 2)\, N_{th}$ used in practice, the synchronism band usually decreases, since, in the circuit with the phase shifter,

* Figures 7a, b are presented for the primary branch of the resonance characteristic.

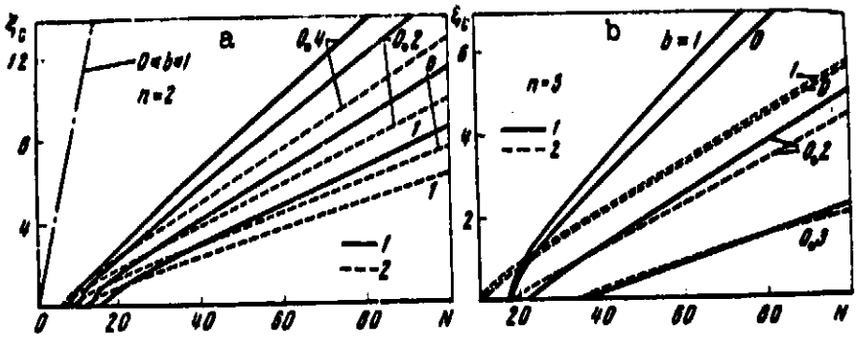


Figure 7. Dependence of DCA synchronism band, expressed in units of detuning ξ_c, on regeneration coefficient N:

a — n = 2.4; B — n = 5; 1 — $\Phi = 0$; 2 — $\Phi = \pi/2$

the synchronism band is determined not by the quadrature component γ_s, but rather by the in-phase component γ_c which, for the indicated values of N, is usually smaller than γ_s.

Thus, in contrast with the assumptions of certain authors [3, 12], we can consider it to be established that the so-called "optimal" phase shifter does not increase the synchronism band. The nature of synchronism band dependence on the external signal amplitude (Figure 8) confirms this conclusion. We see from the curves for $n \geq 4$ the synchronism band is maximal for the external signal amplitude $E = 5\ E_0$.

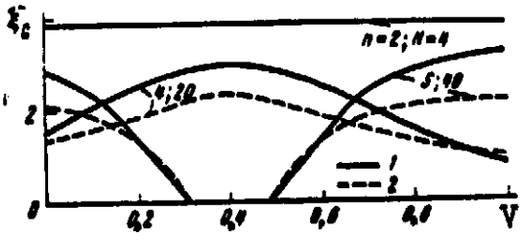


Figure 8. Dependence of synchronism band on external signal amplitude:

1 — $\Phi = 0$; 2 — $\Phi = \pi/2$

The amplitude characteristic, i.e., the dependence of the feedback voltage amplitude on the external signal amplitude for the resonant loop tuning case $\xi = 0$ and $n = 2, 4, 5$; $\Phi = 0$ is shown in Figure 9.

It is convenient to plot the characteristic in the coordinates (a/b, 1/b), since $a/b = (U/E) \cdot (E/2\ E_0) = U/2\ E_0$ and $1/b = E/2\ E_0$.

The values of a and b are determined directly from the EC graphs, drawing a horizontal line at the level 1/N until intersecting the curve γ_c (a) for $\Phi = 0$ (or the curve γ_s for $\Phi = \pi/2$). The amplitude characteristics are nearly linear. In the input amplitude region $1/b = 2.5$, we observe disruption of the synchronism regime for odd n. Depending on the regeneration magnitude, hysteresis in the oscillation onset and termination is possible. The hysteretic part of the characteristics is located to the left of the dash-dot line. The phase shifter does not alter the nature of the curves, but for $\Phi = \pi/2$ they lie considerably higher, since in the case the self-oscillation amplitude a is larger. Marked increase of the oscillation amplitude may cause deterioration of the output voltage form (in the transistorized scheme).

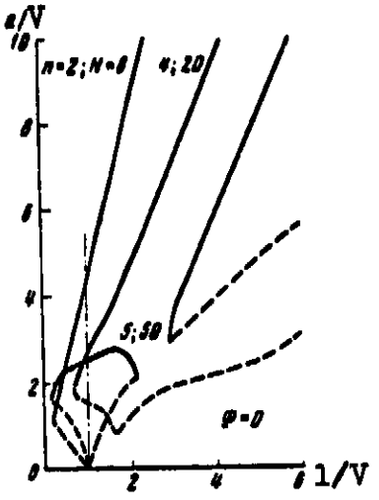


Figure 9. Amplitude characteristic of DCA

Thus, summarizing, we note the following:

1. The subharmonic current amplitude at the FC output is minimal in the absence of phase shifts in the feedback loop, and is maximal when connecting the "optimal" phase shifter into the circuit, which introduces the phase shift $\Phi = \pm \pi/2$, corresponding to the feedback voltage phase $\phi = \pm \pi/2$ n. /102

2. The introduction of the phase shifter ($\Phi = \pi/2$, $n > 2$) may reduce the regeneration threshold magnitude by a factor of 1.5 to 2. Here, the synchronism band decreases in most cases, and the frequency characteristics deteriorate, which leads to increase of the phase instability. Deterioration of the output signal shape is also possible. These negative phenomena are accentuated with increase of the division ratio n. /103

3. The use of the phase shifter for frequency division by a factor of two is not advisable in general because of the absence of any advantages. In those rare cases of FD use when phase stability does not play a significant role and, figuratively speaking, the simple fact of frequency division is sufficient, for $n > 2$ a phase shifter can be introduced into the circuit in order to lower the regeneration threshold. In all other cases of FD use (in devices for synchronizing and frequency tracking in phase-sensitive systems) circuits without a phase shifter can be recommended in order to improve FD phase stability. All that we have said here also applies in full measure to frequency dividers with reactive feedback [9, 10], so that their application is not advisable.

4. In this study, we have determined the optimal external signal amplitude for which the FD has the maximal synchronism band and minimal threshold for $n \geq 4$. We have identified the existence of a region of "critical" input amplitudes where generation of odd subharmonics of multiplicity $n \geq 3$ is not possible. The existence of hysteresis in the onset and termination of the oscillations has been established.

The author wishes to thank Prof. S. I. Evtyanov and V. V. Andreyanov, Cand. Tech. Sci., for their review of the manuscript and valuable comments.

REFERENCES

1. Fitzgerald, J. A. Electronic Eng., Vol. 24, No. 295, 1952, p. 413. /104

2. Korolev, P. G. Trudy VKIAS, No. 58, 1957, p. 137.

3. Il'minskiy, N. Ya. and Ye. G. Loyter. In the collection: Poluprovodnikovye pribory i ikh primeniye (Semiconductor Devices and Their Application). Soviet Radio Press. Moscow, Vol. 5, 1960, p. 255.

4. Andreyev, V. S. and M. Z. Tseytlin. Electrosvyaz', Vol. 4, 1959, p. 23.

5. Dem'yanchenko, A. G. and S. I. Yevtyanov. Radiotekhnika, Vol. 17, No. 10, 1962, p. 25.

6. Bykov, V. L. Transactions of the State Scientific Research Institute of the Ministry of Communications of the USSR, Vol. 2 (34), Vol. 3 (35), 1964.

7. Yevtyanov, S. I. NDVSh (Radiotekhnika i elektronika), Vol. 2, 1958, p. 134.

8. Laine, E. L. Electronic Industries, Vol. 12, 1958, p. 62.

9. Utkin, G. M. NDVSh (Radiotekhnika i elektronika), Vol. 2, 1958, p. 151.

10. Utkin, G. M. Trudy MEI, Vol. 34, 1961, p. 161.

11. Yevtyanov, S. I. Radiotekhnika, Vol. 11, No. 6, 1956, p. 3.

12. Rivkin, I. Kh. Umnozhiteli i deliteli chastoty (Frequency Multipliers and Dividers). Svyaz' Press, 1966.

13. Andreyev, V. S. Radiotekhnika, Vol. 16, No. 9, 1961, p. 60

14. Dem'yanchenko, A. G. and Yu. G. Meshman. Radiotekhnika i elektronika, Vol. 14, No. 11, 1969, p. 1957.

15. Yevtyanov, S. I. Radiotekhnika, Vol. 1, No. 1, 1946, p. 68.

16. Bruevich, A. N. and S. I. Yevtyanov. Approksimatsiya nelineynykh kharakteristik i spektry pri garmonicheskom vozdeystviy (Approximation of Non-Nonlinear Characteristics and Spectrum with Harmonic Input). Soviet Radio Press, Moscow, 1965.

COMB-LINE BANDPASS FILTERS

Ye. A. Vlasov

ABSTRACT. A calculation is made of comb-line microwave filters with respect to specified losses with the aid of a prototype low-frequency filter. We examine a technique for calculating shielded parallel coupled lines with circular cross section conductors. Along with the formulas and graphs, we give an example of the calculation of a comb-line bandpass filter with respect to specified parameters. We examine specific designs of such filters in application to their use in spacecraft onboard antenna and feeder devices.

/104

The radiotechnical system of the modern spacecraft includes several radio receiving and radio transmitting devices which can operate simultaneously in various frequency channels, providing two-way radio communication, transmission of telemetry information, and the conduct of various physical measurements. Because of the comparatively small dimensions, the antennas of these devices are inadequately isolated from one another, which leads to the appearance of mutual interference. For interference suppression, the channels of the devices include bandpass filters, each of which passes the signal band of its device and suppresses the noise signals. In addition, a single antenna is often used for simultaneous operation of several devices on spaced frequencies. Bandpass filters are usually used for combining or separating the signals of these frequencies [1].

Comb-line bandpass filters (BPF) [2, 3] have been widely used in the microwave (SHF) band. Figure 1 shows the electrical schematic of such a filter. It contains n parallel coupled metal bars 1, ..., n of length $l = \lambda/8$ or shorter. The bars are grounded at one end, and loaded by the capicitors C^k at the other end. The end bars 1 and n are connected in parallel with the matching transformers 0 and n + 1, which are also grounded on one side, while the other side is connected to the circuit. Such filters are small, since the bars are shorter than $\lambda/8$, and simple to fabricate. /105

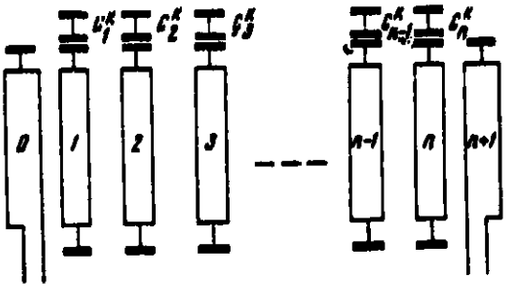


Figure 1. Electrical diagram of comb-line bandpass filter (BPF)

Tuning of the filter is accomplished by the capacitors, which makes it possible to compensate for fabrication errors. The filters have low losses, since the use of dielectric materials is minimized. The first parasitic passband is located no nearer than the fourth harmonic.

Filter frequency characteristic. The frequency characteristic of the ideal bandpass filter is shown in Figure 2. The filter passes a signal without loss in the frequency band $\Delta f = f_p - f_{-p}$, and has attenuation equal to infinity outside this band. However, such a characteristic is physically unrealizable [4]. The real filter has some attenuation L_n in the passband; this attenuation depends on the frequency. Usually it is approximated in the passband in the form of the maximally-flat (Figure 3a) or Chebyshev (Figure 3b) characteristic. Outside the passband, the filter attenuation also depends on the frequency.

In addition to the basic passband, microwave filters have parasitic bands which must be considered in their design. Usually, the

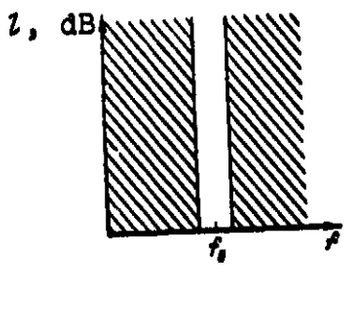


Figure 2. Frequency characteristic of ideal BPF

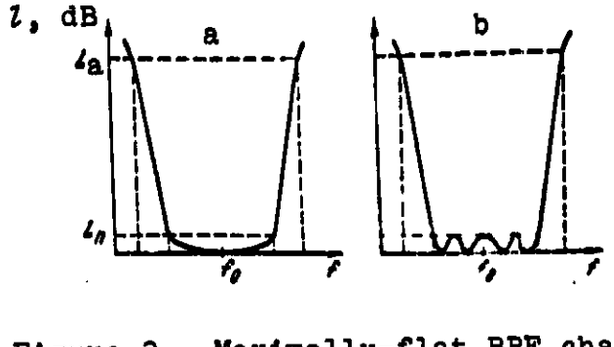


Figure 3. Maximally-flat BPF characteristic (a) and Chebyshev BPF characteristic (b)

filter frequency characteristic is specified for its design: passband center frequency $f_0 = (f_p + f_{-p})/2$, passband relative width $w = \Delta f/f_0 = (f_p - f_{-p})/f_0$, the pulsation amplitude in the passbsnd for the filter with Chebyschev characteristic or the attenuation L_n at edge of the passband for the filter with maximally-flat frequency characteristic, the frequency f_{-a} or f_a outside the passband at which the attenuation L_a must be provided.

Reference [3] presents several graphs for determining the number of resonators of filters with Chebyshev or maximally-flat characteristic. It is convenient to determine the number of resonators for any filter with Chebyshev characteristic from the nomogram of Figure 4 [5].

Example. Let us design a comb-line BPF with Chebyshev characteristic having the following parameters: passband center frequency f_0 = 1223 MHz; passband width w = 0.08; pulsation amplitude in the passband L_n = 0.1 dB; the filter must provide attenuation L_a = 50 dB at the frequency f_a = 1616 MHz; the filter is connected to a line with characteristic impedance Z_A = 50 ohm. We determine the normed frequency variable:

$$\Omega = 2 \cdot \frac{f_a - f_n}{f_p - f_p} = \frac{2}{w}\left(\frac{f_a}{f_n} - 1\right), \qquad (1)$$

in our case, $\Omega = \frac{2}{0,08}\left(\frac{1616}{1223} - 1\right) = 8$.

From the nomogram (Figure 4), we find that the value n = 3 corresponds to the values Ω = 8, L_n = 0.1 dB, L_a = 50 dB. /106

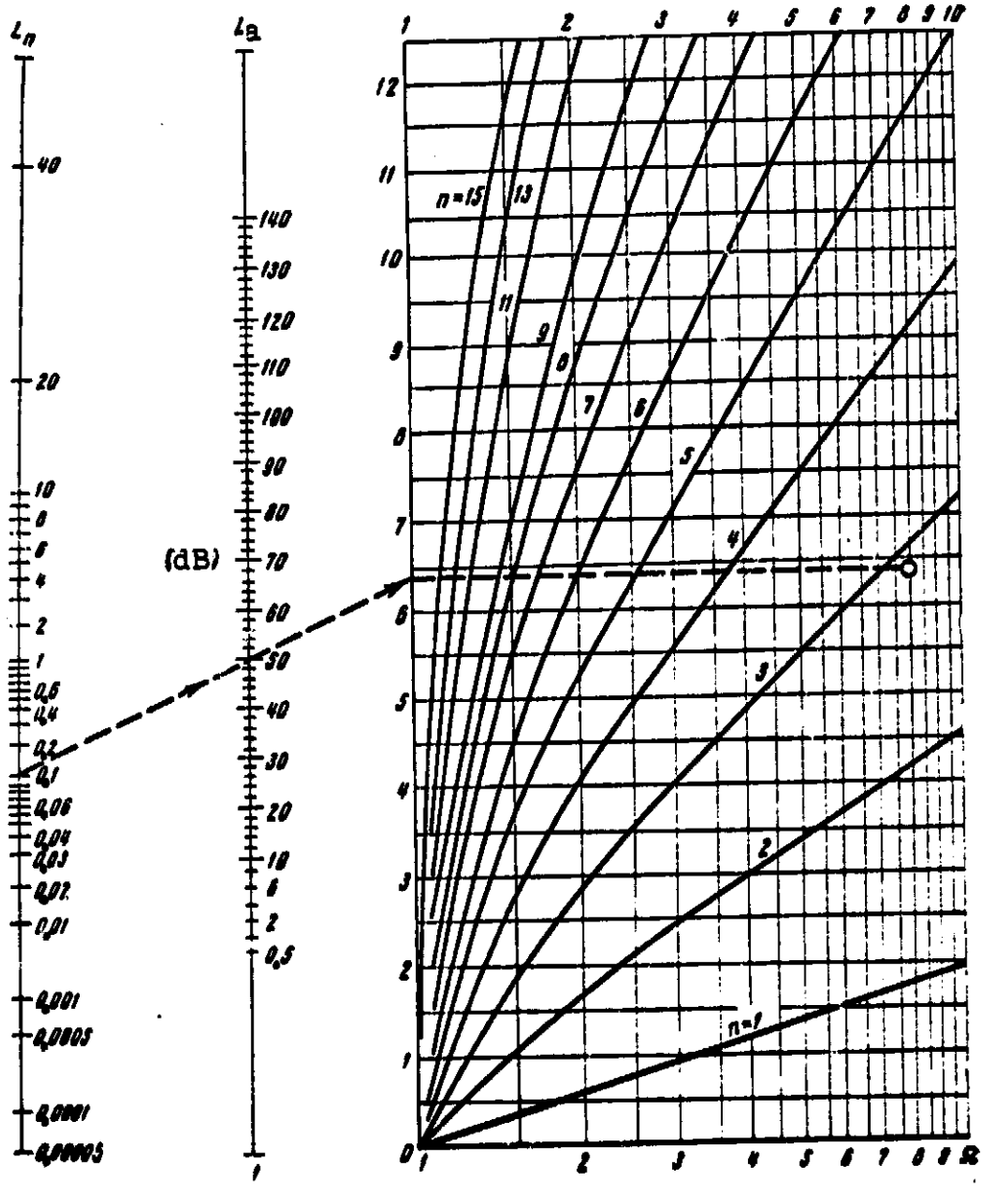


Figure 4. Nomogram for determining number n of resonators for filter with Chebshevskiy characteristic n = 1 - 15

Resonator parameters. Historically, the development of microwave filter theory was preceded by the development of low-frequency filter theory. A design technique was developed for the latter, and extensive tables of values of the individual parameters were compiled. The low-frequency filters were taken as prototypes in developing the technique for microwave filter design. Specifically, the low-frequency filter with lumped elements is usually used as the prototype for the design of comb-line filters [2, 3] (Figure 5).

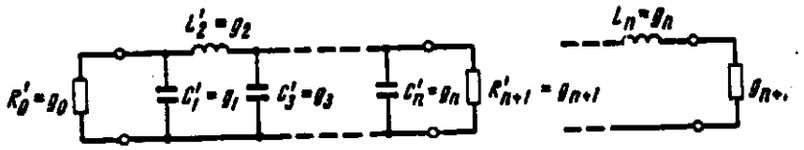


Figure 5. Schematic of ladder filter used as low-frequency prototype

Tables of prototype filter parameters for Chebyshev and maximally-flat characteristics with number of resonators from 2 to 15 are presented in [3, 6]. /107

The microwave comb-line filter resonators are characterized by the characteristic impedance Z_{aj} and electrical length θ. Figure 6 shows coaxial line quality factor as a function of characteristic impedance [3]. The optimal quality is obtained for $\sqrt{\varepsilon_r} \cdot Z_0 = 74$ ohm. With some approximation, this value of the characteristic impedance was taken for comb filters with bars of rectangular section, and yielded quite good results [2 - 4]. Good results are also obtained for the choice $Z_{aj} = 74$ ohm in the case of circular bars. The electrical length of the comb filter resonator bars is usually $\theta = 2\pi l/\lambda \leq \pi/4$.

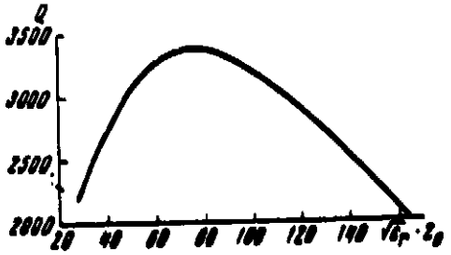


Figure 6. Coaxial line quality factor versus characteristic impedance

Example. For a filter with n = 3 and L_a = 0.1, we have the following values of the elements: g_0 = 1; f_1 = 1.0315; g_2 = 1.1474; g_3 = 1.0315; g_4 = 1.

From the graph (see Figure 6), we select Z_{aj} = 74 ohm, which corresponds to Y_{aj} = 0.0135 mho. It was previously assumed that Z_A = 50 ohm, i.e., Y_A = 0.02 mho.

$$\frac{Y_{aj}}{Y_{A1}} = 0{,}677, \quad \frac{Y_{A1}}{Y_{aj}} = 1{,}48.$$

Determination of partial capacitances. Determination of the partial capacitances is preceded by determination of the parameters which are functions of the electrical length θ:

$$\frac{b_j}{Y_{aj}}, \frac{C_{t1}}{Y_{aj}}, \frac{C_{tn}}{Y_{aj}}, \frac{J_{j,j+1}}{Y_{aj}} \operatorname{tg}\theta \Big|_{j=1+(n-1)}.$$

It is convenient to determine these parameters with the aid of nomograms and graphs. Figure 7 shows the relation:

$$\frac{b_j}{Y_{aj}} = f(\theta) = \frac{\operatorname{ctg}\theta + \theta \operatorname{cosec}^2\theta}{2}. \qquad (2)$$

Figure 8 shows the nomogram for determining

$$\frac{C_{t1}}{Y_{aj}} = \frac{C_{tn}}{Y_{aj}} = \varphi\left(\frac{b_j}{Y_{aj}}\right) = \frac{w}{g_0 g_1 \omega_1'} \frac{b_j}{Y_{aj}} \qquad (3)$$

and

$$\frac{J_{j,j+1}}{Y_{aj}} \operatorname{tg}\theta = \psi\left(\frac{b_j}{Y_{aj}}\right) = \frac{w}{\omega_1' \sqrt{g_j g_{j+1}}} \frac{b_j}{Y_{aj}} \operatorname{tg}\theta \qquad (4)$$

for the parameters of the prototype filter with n = 3 and L_n = 0.1 dB. In the left

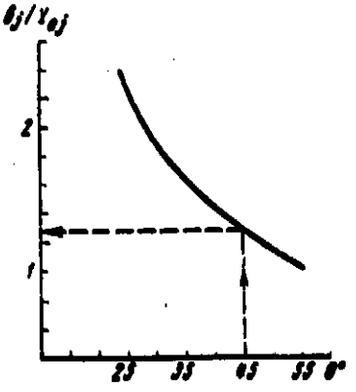


Figure 7. Parameter b_j/Y_{aj} versus resonator electrical length θ

/109

side of the nomogram are plotted curves of the relations:

$$\frac{G_{tL}}{Y_{aj}} = \varphi\left(\frac{b_j}{Y_{aj}}\right) \quad \text{and} \quad \frac{J_{j,j+1}}{Y_{aj}} \operatorname{tg}\theta = \psi\left(\frac{b_j}{Y_{aj}}\right)$$

for w = 0.1. The right side of the nomogram is used to determine the values of these parameters for w = 0.01 - 0.1. Nomograms for any other values of n, L_n, w can be constructed using this same principle.

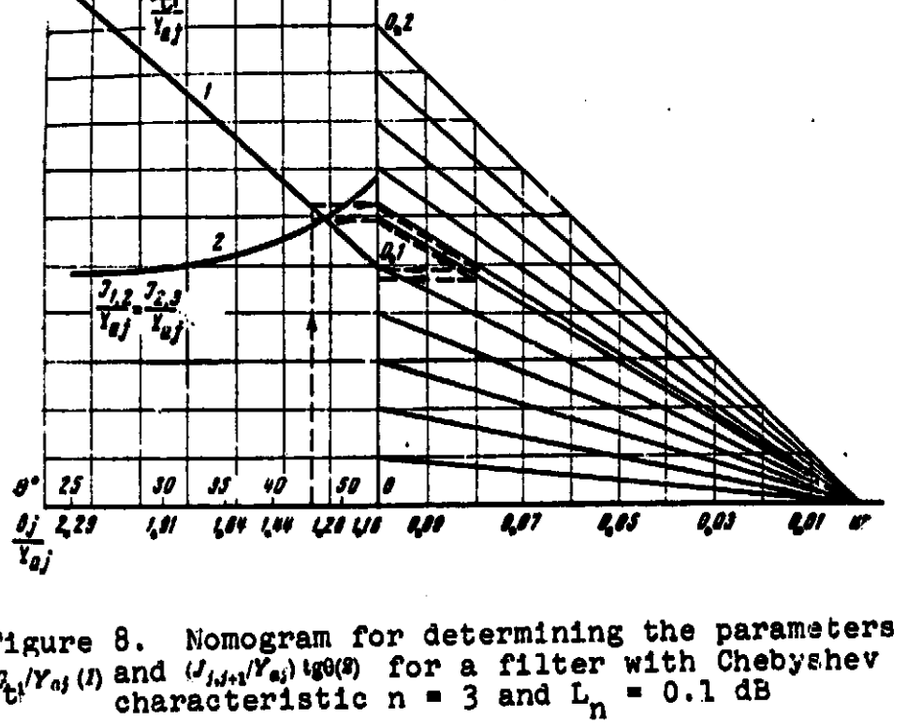


Figure 8. Nomogram for determining the parameters G_{tj}/Y_{aj} (1) and $(J_{j,j+1}/Y_{aj})\operatorname{tg}\theta$(2) for a filter with Chebyshev characteristic n = 3 and L_n = 0.1 dB

In view of the fact that the values obtained with the aid of the nomograms are normed relative to Y_{aj} (normed relative to Y_A in Mattaei [2]), the formulas for determining the capacitances change somewhat:

$$\frac{C_n}{\varepsilon} = \frac{370{,}7 \cdot Y_A}{\sqrt{\varepsilon_r}}\left(1 - \sqrt{\frac{G_{tL}}{Y_{aj}}\frac{Y_{aj}}{Y_A}}\right), \tag{5}$$

$$\frac{C_1}{\varepsilon} = \frac{370{,}7 \cdot Y_{aj}}{\sqrt{\varepsilon_r}}\left(1 - \frac{Y_A}{Y_{aj}} + \frac{G_{tL}}{Y_{aj}} - \frac{J_{1,2}}{Y_{aj}}\operatorname{tg}\theta\right) + \frac{C_0}{\varepsilon}, \tag{6}$$

$$\frac{C_j}{\varepsilon}\Big|_{j=1+n-1} = \frac{376{,}7\cdot Y_{aj}}{\sqrt{\varepsilon_r}}\left(1 - \frac{J_{j-1,j}}{Y_{aj}}\operatorname{tg}\theta - \frac{J_{j,j+1}}{Y_{aj}}\operatorname{tg}\theta\right), \tag{7}$$

$$\frac{C_n}{\varepsilon} = \frac{376{,}7\cdot Y_{aj}}{\sqrt{\varepsilon_r}}\left(1 - \frac{Y_A}{Y_{aj}} + \frac{C_{tn}}{Y_{aj}} - \frac{J_{n-1,n}}{Y_{aj}}\operatorname{tg}\theta\right) + \frac{C_{n+1}}{\varepsilon}, \tag{8}$$

$$\frac{C_{n+1}}{\varepsilon} = \frac{376{,}7\cdot Y_A}{\sqrt{\varepsilon_r}}\left(1 - \sqrt{\frac{C_{tn}}{Y_{aj}}\frac{Y_{aj}}{Y_A}}\right), \tag{9}$$

$$\frac{C_{tn}}{\varepsilon} = \frac{376{,}7\cdot Y_A}{\sqrt{\varepsilon_r}} = \frac{1}{\varepsilon}C_n, \tag{10}$$

$$\frac{C_{j,j+1}}{\varepsilon}\Big|_{j=1+n-1} = \frac{376{,}7\cdot Y_{aj}}{\sqrt{\varepsilon_r}}\left(\frac{J_{j,j+1}}{Y_{aj}}\operatorname{tg}\theta\right), \tag{11}$$

$$\frac{C_{n,n+1}}{\varepsilon_N} = \frac{376{,}7\cdot Y_A}{\sqrt{\varepsilon_n}} - \frac{C_{n+1}}{\varepsilon}. \tag{12}$$

With the aid of the nomogram (Figure 9), we can determine the load capacitor capacitances:

$$C_j^* = \frac{Y_{aj}\operatorname{ctg}\theta}{2\pi\cdot f_0}. \tag{13}$$

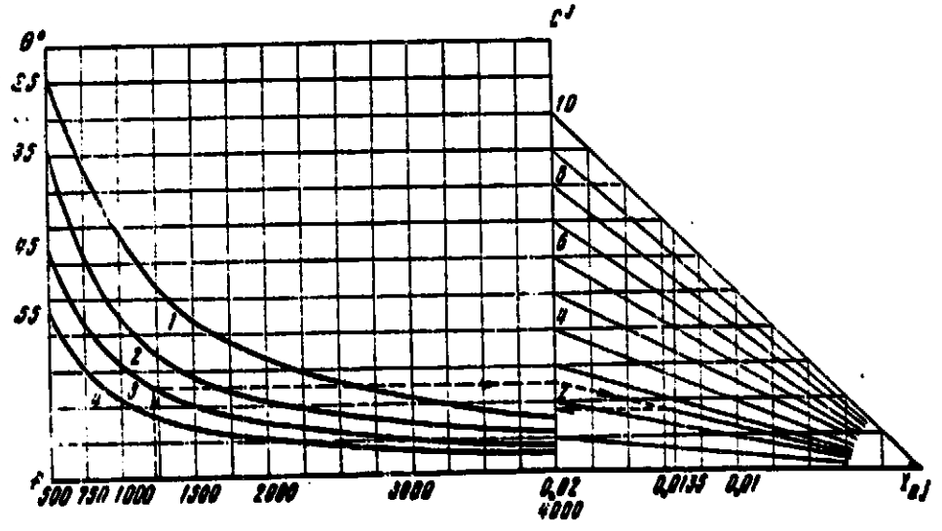


Figure 9. Nomogram for determining capacitances of lumped loading capacitors for several resonator electrical lengths:

1 — θ = 25°; 2 — 35°; 3 — 45°; 4 — 55°

The relation $C_j^k = \phi(f)$ for $Y_{aj} = 0{,}01$ mho for several discrete values of θ is plotted on the left side. The right side of the nomogram serves for determining the quantity C_j^k for arbitrary Y_{aj} in the limits from 0.01 to 0.02 mho.

Example. With the aid of the graph (see Figure 7), we determine b_j/Y_{aj}. For $\theta = \alpha/4 = 45°$, $b_j/Y_{aj} = 1.28$. For the found value of b_j/Y_{aj}, using the nomogram of Figure 9, we determine $\frac{G_{t1}}{Y_{aj}} = \frac{G_{t2}}{Y_{aj}}$, and $\frac{J_{1,2}}{Y_{aj}} \operatorname{tg} 0 = \frac{J_{2,3}}{Y_{aj}} \operatorname{tg} 0$.

To the quantity $b_j/Y_{aj} = 1.28$ corresponds $G_{t1}/Y_{aj} = 0.124$ for $\omega = 0.1$. In order to determine G_{t1}/Y_{aj} for $\omega = 0.08$, we draw from the point $G_{t1}/Y_{aj} = 0.124$ on the vertical axis, a sloping ray to intersect the point 0 of the horizontal w axis. The point of intersection of the ray with the vertical straight line $w = 0.08$ is then projected to the G_{t1}/Y_{aj} axis, and the point of intersection yields the desired value $G_{t1}/Y_{aj} = 0.099$. We can find the quantity $\frac{J_{1,2}}{Y_{aj}} \operatorname{tg} 0$: /110
similarly:

for $w = 0{,}1$ $\quad \frac{J_{1,2}}{Y_{aj}} \operatorname{tg} 0 = 0{,}119$,

for $w = 0{,}08$ $\quad \frac{J_{1,2}}{Y_{aj}} \operatorname{tg} 0 = 0{,}094$.

The bar self-capacitances per unit length have the values:

$$\frac{C_{11}}{\varepsilon} = \frac{376{,}7 \cdot 0{,}02}{1}(1 - \sqrt{0{,}099 \cdot 0{,}677}) = 5{,}58 = \frac{C_{\text{в}}}{\varepsilon},$$

$$\frac{C_{12}}{\varepsilon} = \frac{376{,}7 \cdot 0{,}0135}{1}(1 - 1{,}48 + 0{,}099 - 0{,}094) + 5{,}58 = 3{,}17 = \frac{C_{\text{в}}}{\varepsilon},$$

$$\frac{C_{13}}{\varepsilon} = \frac{376{,}7 \cdot 0{,}0135}{1}(1 - 0{,}094 - 0{,}094) = 4{,}13.$$

The mutual capacitances per unit length are:

$$\frac{C_{0,1}}{\varepsilon} = \frac{C_{n,n+1}}{\varepsilon} = 7{,}5 - 5{,}58 = 1{,}92,$$

$$\frac{C_{1,2}}{\varepsilon} = \frac{C_{n-1,n}}{\varepsilon} = \frac{370{,}7 \cdot 0{,}0135}{1} \cdot 0{,}094 = 0{,}48.$$

From the nomogram of Figure 9, we determine the load condenser capacitances. For f_0 = 1.223 MHz, θ = 45°, with Y_{aj} = 0.02 the capacitance C^k = 2.55 pF, while for Y_{aj} = 0.0135, the capacitance C^k = 1.75 pF.

Determination of filter geometric dimensions. The comb-line filter is built structurally in the form of a system of parallel resonator bars mounted in a single plane, on both sides of which at equal distance from the corresponding bars there are located screening plates. By filter geometric dimensions, we mean the cross section dimensions of the individual resonator bars, the distance between adjacent bars, and the distance between the screening plates. Usually, the bars are circular (Figure 10) or rectangular (Figure 11) in cross section. The self and mutual capacitances of the resonator bars are shown in the figures.

Other conditions being the same, the structures with circular bars are somewhat simpler to construct. We shall examine the technique for calculating the cross section geometric dimensions of a filter with circular bars. The filter structure is broken down into a number of elementary cells equal to

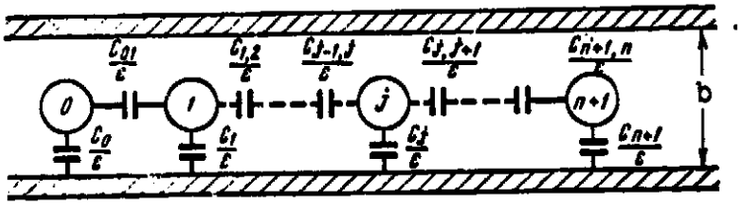


Figure 10. Cross section of filter with circular bars

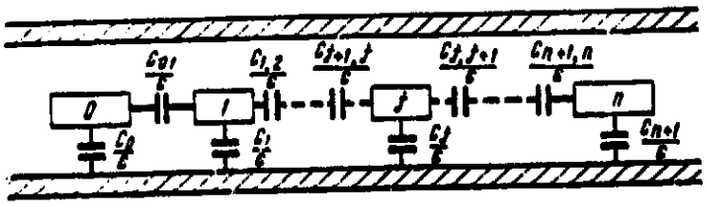


Figure 11. Cross section of filter with rectangular bars

the number of bars. Each cell is a T-shaped combination of three condensers, one of which represents the self-capacitance per unit length of the corresponding bar, and the other two represent the coupling capacitances of this bar with the neighboring bars.

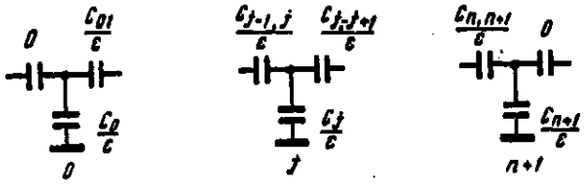


Figure 12. Elementary filter cells

Figure 12 shows the filter elementary cells. The unilaterally loaded cells 0 and n++ correspond to unilateral matching transformers at the filter input and output.

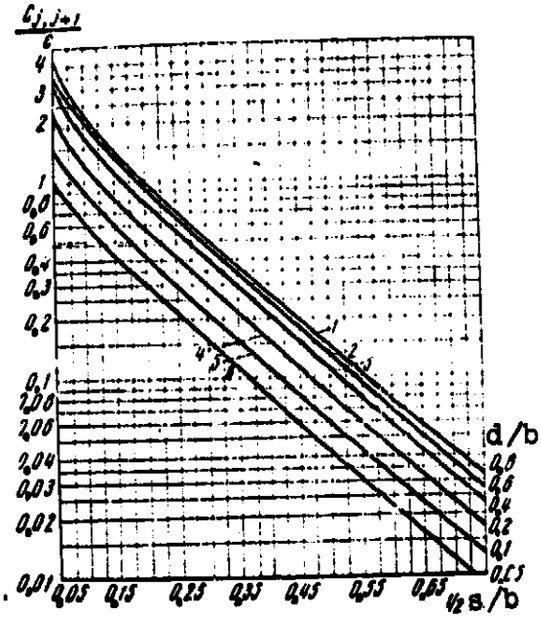


Figure 13. Relationship between $1/2\ (C_{j,j+1}/\varepsilon)$ and $1/2\ (S/b)$ for several fixed values of d/b: 1 — d/b = 0.8; 2 — 0.6; 3 — 0.4; 4 — 0.2; 5 — 0.1; 6 — 0.5

The cells 1 - j - n correspond to the bilaterally loaded resonator bars. With the aid of Figures 13 - 15, for each j^{th} elementary cell we determine the relative bar diameter d_j/b, and half the relative distances from the bar to the two adjacent bars:

$$\frac{1}{2}\frac{S_{j-1,j}}{b} \quad \text{and} \quad \frac{1}{2}\frac{S_{j,j+1}}{b}.$$

Figure 13 shows the dependence of the mutual capacitance $C_{j,j+1}/\varepsilon$ between bars on half the distance between bars $1/2\ S_{j,j+1}/b$ for several values of the relative bar diameter d/b of the j^{th} cell.

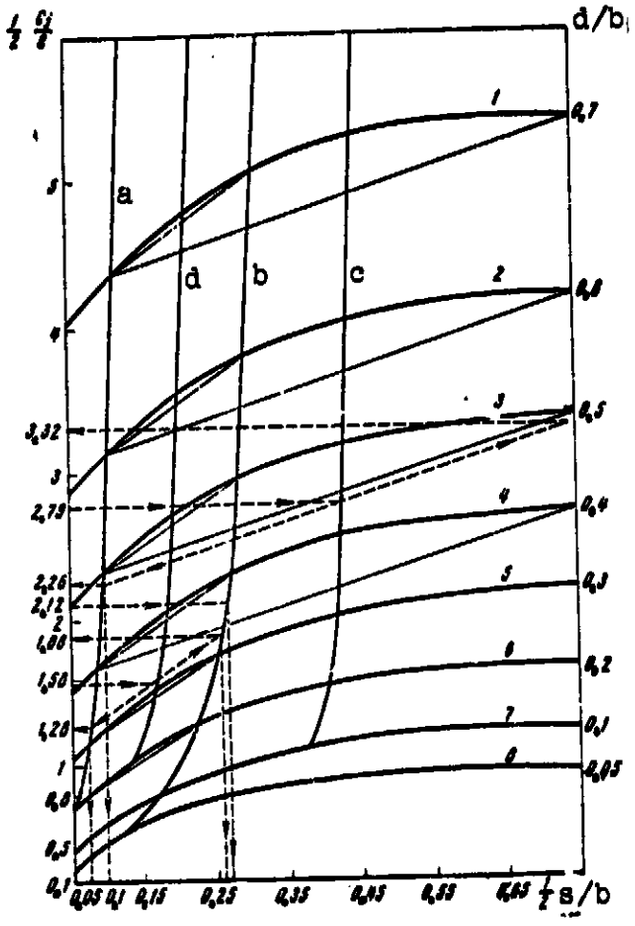


Figure 14. Relationship between 1/2 (C_j/ε) and 1/2 (S/b) for several fixed values of d/b:

1 — d/b = 0.7; 2 — 0.6; 3 — 0.5; 4 — 0.4; 5 — 0.3; 6 — 0.2; 7 — 0.1; 8 — 0.05

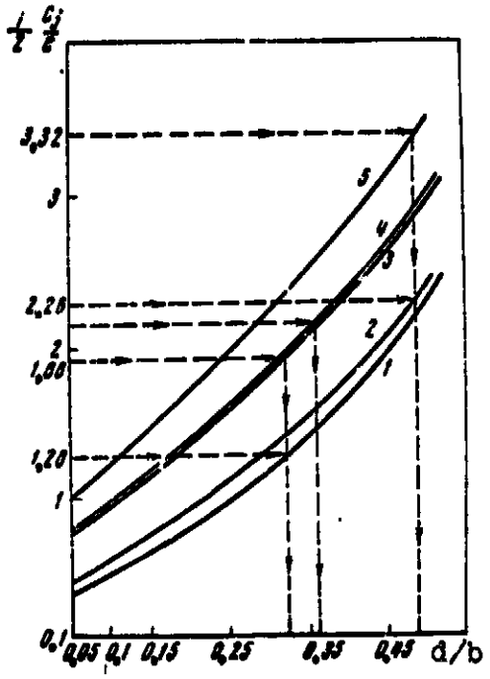


Figure 15. Relationship between 1/2 (C_j/ε) and d/b for several fixed values of 1/2 (S/b):

1 — 1/2 (S/b) = 0.07; 2 — 0.1; 3 — 0.26; 4 — 0.27; 5 — 0.75

Figure 14 shows the dependence of half the self-capacitance 1/2 C_j/ε of the j^{th} bar on half the distance between the bars 1/2 $S_{j,\,j+1}/b$ for several values of d/b. /111

In order to determine the dimensions of the bilaterally loaded element, we draw on Figure 13, horizontal lines corresponding to the element mutual capacitances $C_{j-1,j}/\varepsilon$ and $C_{j,\,j+1}/\varepsilon$. The coordinates of the points of intersection of these lines with the lines (1) - (5) of constant d/b values are transferred from Figure 13 to Figure 14, and from them we plot the curves for $C_{j-1,\,j}/\varepsilon$ and $C_{j,\,j+1}/\varepsilon$. The sought values of 1/2 $S_{j-1,j}/b$ and 1/2 $S_{j,j+1}/b$ are the horizontal

coordinates of the points of intersection of the constructed curves with the curve corresponding to the sought d_j/b value. On the other hand, the sum of the vertical coordinates of these points is equal to the resonator self-capacitance. In order to find the indicated points of intersection, we construct an auxiliary curve through the midpoint of the chords joining the points of intersection of the curves of constant d/b values with the constant mutual capacitance curves. On the graph, we draw a horizontal line corresponding to half the rod self-capacitance to intersect the auxiliary curve. Through the point of intersection we draw a line parallel to the previously mentioned chords. The points of intersection of this line with the curves $C_{j-1,j}/\varepsilon$ and $D_{j,\ j+1}/\varepsilon$ lie on the sought d_j/b curve. The projections of these points on the abscissa axis yield the sought values of $1/2\ S_{j-1,j}/b$ and $1/2\ S_{j,\ j+1}/b$.

The sought values of d_j/b are found with the aid of Figure 15. On the vertical axis, we ley off the partial self-capacitances $1/2\ C_j/\varepsilon$ of the bars. Along the horizontal axis, we lay off the values of d/b. To find the sought d/b value, we draw in Figure 14, vertical lines corresponding to the found values of $C_{j-1,j}/\varepsilon$, and $C_{j,\ j+1}/\varepsilon$ find their points of intersection with the curves of constant d/b values, and transfer them to Figure 15. Through these points in Figure 15, we draw curves of the found values of $1/2\ S_{j-1,j}/b$ and $1/2\ S_{j,j+1}/b$. /113
On the vertical axis of the figure, we lay off the previously found self-capacitance half-values, which in sum yield the self-capacitance C_j/ε of the j^{th} bar. Through these points we draw horizontal lines to intersect the corresponding $1/2\ S_{j-1,j}/b$ and $1/2\ S_{j,j+1}/b$ curves. The points of intersection are projected to the horizontal axis of the figure to a single point corresponding to the sought d_j/b value.

This technique applies to the bilaterally loaded bars.

The unilaterally loaded bars, specifically the matching transformers at the filter input and output, have some finite mutual coupling capacitance $C_{0,1}/\varepsilon$ with the neighboring bar of the filter structure on one side, and the coupling capacitance $C_{0,0}/\varepsilon = 0$ on the other side. Theoretically, zero coupling capacitance is obtained with $S_{0,0}/b = \infty$; in practice, good results are obtained with $S_{0,0}/b \geq 1.5$. In calculating the cell with unilaterally loaded bar, we take the vertical line 1/2 S/b = 0.75, as the line of zero capacitance.

Example. We determined previously the self and mutual capacitances of the resonator bars. The subject filter can be represented in the form of five cells, and, in view of symmetry, cells 1 and 5, 2 and 4 are pairwise analogous. In Figure 13, we draw two horizontal lines corresponding to the previously determined values of $C_{0,1}/\varepsilon = C_{3,4}/\varepsilon = 1.92$ and $C_{1,2}/\varepsilon = C_{2,3}/\varepsilon = 0.48$, to intersect the lines of constant d/b values (curves 1 - 5). We transfer the points of intersection to Figure 14, and draw through them curves corresponding to these capacitances. The first filter cell is a unilaterally loaded bar with the partial self-capacitance $C_0/\varepsilon = 5.58$, and the partial coupling capacitances $C_{0,0}/\varepsilon = 0$, $C_{0,1}/\varepsilon = 1.92$. In Figure 14, we make an additional construction: the points of intersection of each of the constant d/b lines (1) - (8) with the curve a $C_{0,1}/\varepsilon = 1.92$ and the line 1/2 S/b = 0.75 are joined by chords. Through the chord midpoint, we draw the auxiliary curve b. Through the point on the vertical axis of the figure corresponding to the value $1/2\ C_0/\varepsilon = 5.58/2 = 2.79$, we draw a horizontal line to intersect the constructed auxiliary curve b. Through this point of intersection, we draw a line parallel to the previously constructed chords to intersect the lines a ($C_{0,1}/\varepsilon = 1.92$) and 1/2 S/b = 0.75. The points of intersection have the coordinates (2.26; 0.1) and (3.32; 0.75). The sum of the vertical coordinates 2.26 + 3.32 = 5.58 = C_0/ε, which was to be proved, i.e., the found points lie on the curve of the sought d_0/b value. The

horizontal coordinates yield the half-distances:

$$\frac{1}{2}\frac{S_{0,0}}{b}=0{,}75,\quad \frac{1}{2}\frac{S_{0,1}}{b}=0{,}1.$$

To find the dimension d_0/b, we find in Figure 14 the points of intersection of the vertical lines corresponding to the values $1/2\ S_{0,0}/b = 0.75$ and $1/2\ S_{0,1}/b = 0.1$ with the curves of constant d/b values, and the found points are transferred to Figure 15. In /114 this figure, half the bar self-capacitances are plotted along the vertical axis, and the d/b values are plotted along the horizontal axis. Using the transferred points, we draw curves of constant $1/2\ S/b = 0.1$ (curve 2) and $1/2\ S/b = 0.75$ (curve 5) values. Then we draw horizontal lines of the previously found values of $1/2\ Cb/\varepsilon = 3.32$ to intersect the curve 5, and $1/2\ C_0/\varepsilon = 2.26$ to intersect the curve 2. The points of intersection are projected onto the horizontal axis to a single point $d_0/b = 0.49$. In view of symmetry, cell 5 has the same dimensions.

Cell 2 has the coupling capacitances $C_{0,1}/\varepsilon = 1.92$; $C_{1,2}/\varepsilon = 0.48$, curve b. The auxiliary construction is made similarly, and through the chord midpoints we draw the auxiliary curve d; through the point of intersection of the auxiliary curve d with the lines $1/2\ C_1/\varepsilon = 3.17/2 = 1.585$, a straight line is drawn parallel to the chords. The points of intersection of the parallel line with the curves a — $C_{0,1}/\varepsilon = 1.92$, and b — $C_{1,2}/\varepsilon = 0.48$ have the coordinates (1.28; 0.07) and (1.88; 0.26). Hence the half-distances to the neighboring bars are:

$$^1/_2\ S_{0,1}/b=0{,}07,\quad ^1/_2\ S_{1,2}/b=0{,}26.$$

In order to determine the bar diameter d_1/b, the points of intersection of the vertical lines $1/2\ S/b = 0.07$ and $1/2\ S/b = 0.26$ with the lines of constant d/b values (1 - 8) (Figure 14), are transferred to Figure 15, and through them we draw the curves 1 and 3

corresponding to these values 1/2 S/b = 0.07 and 1/2 S/b = 0.26, and find the points of intersection of these curves with the horizontal lines 1/2 C_1/ε = 1.88 and 1/2 C_1/ε = 1.28, respectively. The found points, projected onto the horizontal axis, yield the sought value d_1/b = 0.32. Cell 4 has similar dimensions.

Cell 3 has the same coupling capacitances $C_{1,2}/\varepsilon = C_{2,3}/\varepsilon = 0.48$. which simplifies the calculation: in Figure 14 we draw the horizontal line 1/2 C_2/ε = 4.13/2 = 2.065 to intersect the curve b ($C_{1,2}/\varepsilon$ = 0.48). The horizontal projection of the point of intersection yields the sought value 1/2 $S_{1,2}/b$ = 1/2, $S_{2,3}/b$ = 0.27. The points of intersection of the vertical 1/2 S/b = 0.27 with the lines (1 - 8) of constant d/b values (see Figure 14) are transferred to Figure 15, and through them we draw the corresponding curve 4. The horizontal coordinate of the point of intersection of this curve with the horizontal line 1/2 C_2/ε = 2.065 yields the sought value d_2/b = 0.36

In determining the filter cross section dimensions, the relative distance between the neighboring bars is obtained by summing like half-distances found for each of the bars. The basic dimensions of the filter are summarized in the following table.

Rod No.	0	1	2	3	4
Cell	1	2	3	4	5
$\frac{d}{b}$	$\frac{d_1}{b}=0{,}49$	$\frac{d_1}{b}=0{,}32$	$\frac{d_2}{b}=0{,}36$	$\frac{d_2}{b}=0{,}32$	$\frac{d_1}{b}=0{,}49$
$\frac{S}{b}$	$\frac{S_{0,1}}{b}=1{,}5$	$\frac{S_{0,1}}{b}=0{,}17$	$\frac{S_{1,2}}{b}=0{,}53$	$\frac{S_{2,3}}{b}=0{,}53$	$\frac{S_{3,4}}{b}=0{,}17$	$\frac{S_{4,5}}{b}=1{,}5$

Note: Commas represent decimal points.

Figure 16 shows the filter cross section.

Construction of lumped loading condensers. The filter resonator loading condenser capacitance can be determined with the aid of the nomogram of Figure 9. In the frequency range in question, the

condenser capacitance for the various filters does not exceed 10 pF. Usually such a capacitance is provided by a flat condenser with two plates, one of which is a metal capacitive plate mounted on the end of the resonator bar, and the other is a corresponding metal block which is in electrical contact with the upper and lower metal plates (Figure 17). Usually, these metal blocks are made movable. This makes it possible to regulate loading condenser capacitance in the tuning process, thereby compensating for any possible small deviations in the filter dimensions.

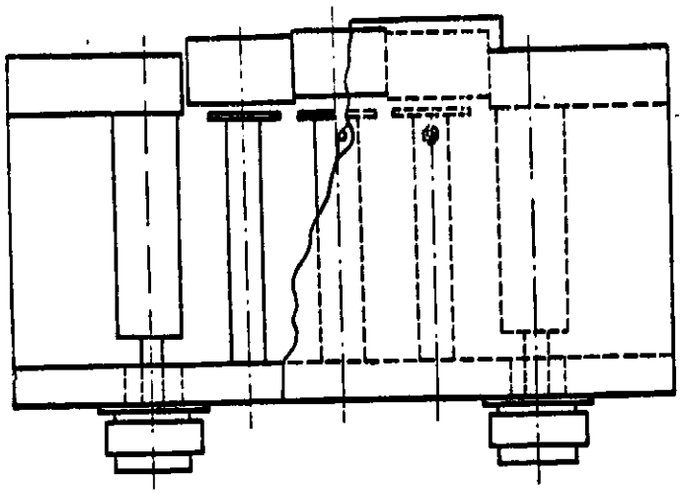


Figure 16. Comb-line filter (top view)

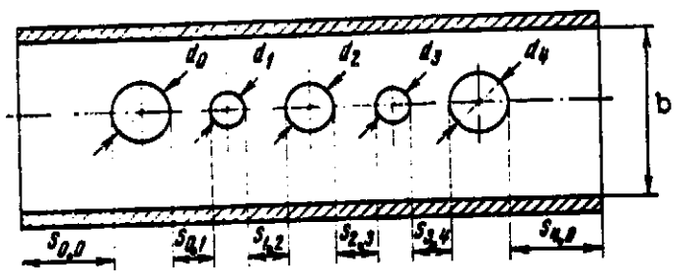


Figure 17. Cross section of comb-line filter with circular bars, calculated in the example

Figure 18 shows a comb filter loading condenser unit [2]. In this unit, tuning of the loading condenser is accomplished in two steps: rough tuning is first accomplished by displacing the metal block, after which the position of the block is fixed, and fine filter tuning is accomplished with the aid of trimming screws. In other words, this process can be termed two-step tuning.

Figure 19 shows a comb filter loading condenser unit which permits accomplishing one-step tuning [9]. Here, a system of threaded joints, including a screw mounted on a block and a screw mounted on a special bracket (yoke), which in turn is rigidly

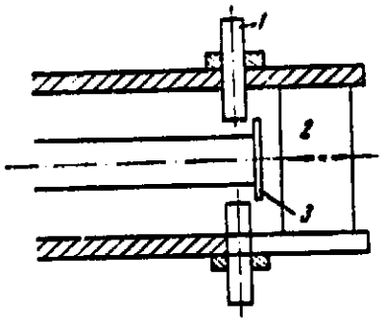


Figure 18. Resonator bar loading condenser unit with "two-step" tuning:

1 — trimming screws; 2 — trimming block; 3 — capacitive plate

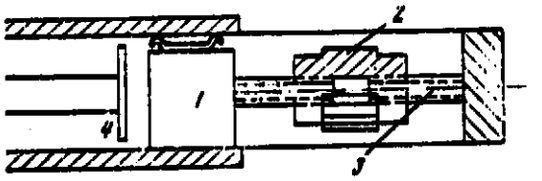
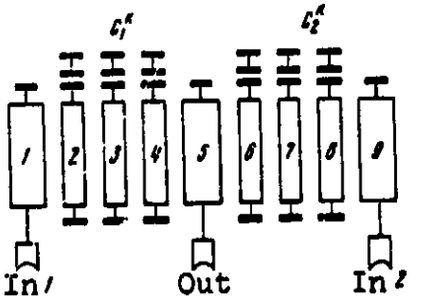


Figure 19. Resonator bar loading condenser unit with "single-step" tuning:

1 — trimmer block; 2 — sleeve; 3 — screws; 4 — capacitive plate

attached to the filter case, is used to displace the block with a small step and fix its position. The screws are made movable relative to one another by means of a metal sleeve with internal thread. Mobility of the entire unit is achieved by making the screws with different pitch or different direction of the threading. As the sleeve is rotated, the metal block displaces, and the displacement step is equal to the sum of the thread pitches when the threads are made in opposite directions, and equal to the difference of the pitches when the screws have threads in the same direction.



/116

Figure 20. Electrical circuit of filter for frequency summation with bandpass elements

Frequency separation using comb-line bandpass filters. Bandpass filters have found wide application as components of circuits for frequency division and summation. Figure 20 shows the electrical circuit of such a filter [10]. The filter has nine bars. Bars 2, 3, 4 and 6, 7, 8 are bandpass element resonators, and bars 1, 5, 9 are matching transformers. The element including resonators 2, 3, 4 is designed to pass a signal with frequency f_1; the element including

resonators 6, 7, 8 is designed to pass a signal with frequency f_2. In this filter, the transformer 5 is loaded on both sides. Here, it is important that the mutual coupling capacitances of transformer 5 with bars 4 and 6 not vary along its length, i.e., that the parallel coupling segment lengths be the same on both sides. This implies that the transformer 5 and resonators 4 and 6 associated with it must have the same geometric length.

As we have noted previously, the resonator bars in the comb-line filter may have electrical length $\theta \leq \pi/4$. This possibility permits making the bars of the subject separation filter of identical geometric length l. The bar electrical length is determined by the geometric length and operating frequency:

$$\theta_1 = \frac{2\pi}{\lambda_1} l_1, \tag{14}$$

$$\theta_2 = \frac{2\pi}{\lambda_2} l_2, \tag{15}$$

where $\lambda = C/2\pi f$ is the wavelength. Since $l_1 = l_2$, the bar electrical /117 lengths are connected with one another by the relation:

$$\frac{\theta_1}{\theta_2} = \frac{\lambda_2}{\lambda_1} = \frac{f_1}{f_2}. \tag{16}$$

In designing the filter, we first select the resonator bar electrical length of one of the bandpass elements, determine the bar geometric length, and use the above relation to determine the electrical length of the bars of the second element. The further calculation is analogous to that of conventional comb-line bandpass elements. The results of an experiment with a separation filter designed using this technique for $\theta_1/\theta_2 = 0.8$ showed good agreement between the characteristics obtained and those assumed in calculating the bandpass elements.

The author wishes to thank T. V. Potapova and N. I. Alekseeva for their assistance in obtaining and analyzing the data presented herein. The author also wishes to thank E. G. Vlostovskiy and L. G. Maloratskiy for their review of individual sections of the manuscript and valuable comments.

REFERENCES

1. Collection: "Apparatura dlya kosmicheskikh issledovaniy". Nauka Press, Moscow, 1972.

2. Matthaei, G. L. Microwave Journal, August, 1963, p. 82.

3. Mattaey, G. L., L. Young and E. M. T. Jones. Microwave Filters, Impedance Matching Networks, and Coupling Structures. Svyaz' Press, Moscow, 1971.

4. Fel'dshteyn, A. L. and L. R. Yavich. Sintez lineynykh chetyrekhpolosnikov i vos'mipolosnikov na SVCh (Synthesis of Linear 4-Terminal and 8-Terminal Microwave Networks). Svyaz' Press, Moscow, 1971.

5. Zverev, A. and S. Kenneth. Electronic Eng., Nov., 1967, p. 44.

6. Fel'dshteyn, A. L., L. R. Yavich and V. P. Smirnov. Spravochnik po elementam volnovodnoy tekhniki (Handbook on Waveguide Elements). Soviet Radio Press, 1967.

7. Cristal, E. G. IEEE Trans. on Microwave Theory and Techn., MTT-12, Vol. 4, 1964, p. 428.

8. Nicholson, B. F. Radio and Electronic Eng., July, 1967, p. 39.

9. Vlasov, Ye. A. Author's Certificate No. 233802, Bulletin of Inventions, No. 3, 1969.

10. Vlasov, Ye. A. Author's Certificate No. 325654, Bulletin of Inventions, No. 3, 1972.

Translated for National Aeronautics and Space Aministration under contract No. NASw 2483, by SCITRAN, P. O. Box 5456, Santa Barbara, California, 93108